DOCUMENT RESUME

ED 206 633 TM 810 474

AUTHOR	Angoff, William H.; Schrader, William B.
TITLE	A Study of Alternative Methods for Equating Rights Scores to Formula Scores.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-81-8
PUB DATE	May 81
NOTE	167p.
EDRS PRICE	MF01/PC07 Plus Postage.
DESCRIPTORS	Academic Ability; College Entrance Examinations; *Equated Scores; Guessing (Tests); Higher Education; *Response Style (Tests); Science Tests; Scoring; *Scoring Formulas; Secondary Education; *Testing Problems; Verbal Tests
IDENTIFIERS	College Board Achievement Tests; Graduate Management Admission Test; Invariance Principle; Scholastic Aptitude Test

ABSTRACT

The purpose of this study was to determine whether it would be possible to equate rights-scored to formula-scored tests without causing a discontinuity in the meaning of the score scale. Several other subsidiary studies--of the characteristics of the two scoring methods, of nonresponse and guessing, and of reliability and parallelism--were also undertaken. The study was conducted in two phases: (1) of two forms of the verbal section of the Scholastic Aptitude Test and one form of the College Board Chemistry Achievement Test; and (2) of operational and experimental subtests of the Graduate Management Admission Test. It was found that the data of this study support the hypothesis that formula scores for tests administered with rights directions are directly comparable to formula scores for the same tests administered with formula directions. Thus, the directions under which a test is administered can be changed without serious concern that a discontinuity in the score scale will result. (Author/GK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED206633

RR-81-8

RESEARCH REPORT

A STUDY OF ALTERNATIVE METHODS FOR EQUATING RIGHTS SCORES TO FORMULA SCORES

William H. Angoff
and
William B. Schrader

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

TM 810 474

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

H. Wiedenmiller

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Educational Testing Service
Princeton, New Jersey
May 1981

A Study of Alternative Methods for Equating Rights Scores to Formula Scores

William H. Angoff
and
William B. Schrader

Educational Testing Service
May 1981

Copyright © 1981. Educational Testing Service. All rights reserved.

A Study of Alternative Methods for Equating
Rights Scores to Formula Scores

Abstract

The principal purpose of this study was to determine whether it would be possible to equate Rights-scored tests to Formula-scored tests without causing a discontinuity in the meaning of the score scale. Several other subsidiary studies--of the characteristics of the two scoring methods, of nonresponse and guessing, and of reliability and parallelism--were also undertaken. The study was conducted in two phases: 1) of two forms of the SAT-verbal and one form of the College Board Chemistry Achievement Test, based on data from a special experimental test administration; and 2) of operational and experimental subtests of the Graduate Management Admission Test, based on data from the (regular) October 1980 administration of the GMAT to applicants for business school. Several outcomes of the study were observed that will be useful for the understanding of some of the issues of Rights and Formula scoring. In addition, it was found that the data of this study support the hypothesis that Formula scores for tests administered with Rights directions are directly comparable to Formula scores for the same tests administered with Formula directions. Thus, the directions under which a test is administered can be changed without serious concern that a discontinuity in the score scale will result.

Acknowledgments

The authors gratefully acknowledge the contributions of Paul W. Holland, Frederic M. Lord, and E. Elizabeth Stewart, who advised us on the analysis and interpretation of the data; of Mark Barton, who was responsible for the statistical calculations; and of Doris Conway, who provided valuable assistance throughout the study.

William H. Angoff
William B. Schrader

A STUDY OF ALTERNATIVE METHODS FOR EQUATING RIGHTS SCORES TO FORMULA SCORES

Contents

Issues in Rights vs Formula Scoring 1

A Study of College Board SAT-verbal and Chemistry Tests, Based on Special Test Administrations 16

Questions Addressed by the Study. 16

Study Design 19

Sample Characteristics 24

Test Administration 30

Results 34

Effect of Directions on Mean Formula Scores: Method of Analysis 34

Effect of Directions on Mean Formula Scores: Findings . . . 36

Effect of Directions on Nonresponse 46

Effect of Directions on Guessing 56

Effect of Directions and Scoring on Reliability and Parallelism 63

Effect of Directions on Equating: Method of Analysis . . . 69

Effect of Directions on Equating: Findings 77

Summary and Conclusions 87

A Study of the Graduate Management Admission Test, Based on Program Data 91

Questions Addressed by the Study 91

Study Design 94

Sample Characteristics 99

Results 100

Effect of Directions on Mean Formula Scores 100

Effect of Directions on Nonresponse 102

Effect of Directions on Guessing 117

Effect of Directions and Scoring on Reliability and Parallelism 125

Effect of Directions on Score Equating: Method of Analysis 131

Effect of Directions on Score Equating: Findings 135

Summary and Conclusions 139

Implications of Findings 142

8

ISSUES IN RIGHTS VS FORMULA SCORING

As is well known, the controversy over the Formula-scoring vs Rights-scoring issue has continued without much loss of force for more than fifty years, and discursive articles and reports of empirical studies appear in the literature on this and related subjects with no less frequency today than in earlier years. The considerations that have persuaded the writers on this topic to adopt one position or the other have ranged widely from issues having to do with the reliability and validity of the scores through the practicalities of administering a testing program under conditions understandable and acceptable to the examinees, to issues of special tactics, ethics, and morality in the taking of tests, differences in personality among examinees and their effects on the (cognitive) test scores, and considerations of equity and fairness to all examinees. Beyond the matter of the particular choice of the administration-and-scoring system is the further complexity of changing from one to the other without causing a discontinuity in the meaning of the scale of scores, a scale that is intended to have the same meaning independently of the form of the test administered, the time and place of administration, or the nature of the examinees tested.

In order that there be no doubt about what is intended in the use of terms and expressions in discussing Formula-vs-Rights issues, it should be understood that answer sheets completed and submitted by examinees are appropriately scored Rights, a simple count of the number of items answered correctly, if the examinees have been informed, prior to their taking the test, that their answer sheets will be scored in this way. This information would normally be accompanied by advice to the examinees that, since there is no penalty for incorrect responses, they should not omit any items in the test. Such words of advice, or directions, may be further described as permissive instructions, in the sense that examinees are encouraged to guess. In contrast to Rights scoring, Formula scoring procedures are those in which an explicit penalty for incorrect responses is used, typically of the form, F (Formula) = R - $\frac{W}{k-1}$, where R and W are the counts of right and wrong responses, respectively, and k = the number of choices per item. The Formula score is not ordinarily intended as a punitive device. The intent of the formula is to yield a net score of zero, on the average, for the aggregate of all items that the examinees mark purely at random. It is understood here too that prior to taking a Formula-scored test the examinee will be informed that this will be the mode of scoring, and he or she will be cautioned that pure guessing is risky and could lower the score. Often added to these instructions is the advice that urges the examinees to eliminate those options that they know are clearly incorrect, and to guess from among the others. Such instructions and advice may be interpreted as restrictive instructions, in the sense that examinees are discouraged from guessing.

It has been pointed out (Diamond and Evans, 1973; Ebel, 1965; Stanley, 1954) that Rights scores and Formula scores are very highly correlated--when all examinees answer all items, the correlation is perfect--and therefore, it might be inferred, it matters very little whether the papers are scored one way or the other. It is indeed true that, given a set of completed answer sheets, the two types of scores are highly correlated. But, clearly, this correlation is spurious; the two scores are based on the same set of responses to the same set of items. What is different is simply the method of dealing with those responses. Further, what is not recognized in making this inference is that in any operational test the method of scoring must go hand in hand with the instructions given to the examinees, and that the test-makers are not at liberty, once they have given the examinees instructions to guess, or not to guess, to score the answer sheets otherwise (Cureton, 1966; Davis, 1967). It may be assumed--and there are sufficient data to support this assumption--that examinees adopt a strategy for guessing, or not guessing, taking the prior information and advice in good faith and assuming that their papers will in fact be scored as they were led to believe they would be. Therefore, the scoring for each individual tested must be performed by the method specified in the directions under which the test is administered.

Rights vs Formula: Pros and Cons

As has been implied at the beginning of this paper, the literature on the issue of Formula scoring and Rights scoring is voluminous, and several excellent reviews of the literature on the topic are available, including extensive reviews by Abu-Sayf (1979) and Diamond and Evans (1973), and a brief overview by Thorndike (1971, 59-61). Consequently, no effort will be made in this paper to conduct an exhaustive review or to evaluate the present state of opinion regarding the various questions regarding Rights and Formula scoring. Instead, a brief summary is made of opinions and findings that bear on the issues as they relate to the conduct of extensive testing programs.

The principal arguments that have been advanced in support of Rights-instructions-and-scoring and in opposition to Formula-instructions-and-scoring are as follows:

1. Rights instructions and advice are much easier for examinees to understand and to follow. Formula advice requires examinees to evaluate their knowledge of the content of the item--e.g., to know when one or more incorrect options can be eliminated--in deciding whether or not to guess.

2. Many examinees fail to understand the logic of the formula, and experience some anxiety about the risks they are asked to take.

3. As a consequence of the foregoing, it is very difficult to write directions to accompany a Formula-scored test that are short, clear, and understandable.

4. It is probably more difficult to attain virtually error-free scoring when Formula-scoring is used than when Rights-scoring is used.

5. Because of the considerations in (1), (2), and (4) above, some investigators claim that a Formula-scored test introduces irrelevant sources of error variance into the test, stemming from differences among examinees with respect to: their ability to understand Formula-score directions and take optimal action (Abu-Sayf, 1979); their levels of confidence in assessing their own degree of knowledge (Slakter, 1968b); their willingness to take a risk (Sherriffs and Boomer, 1954; Slakter, 1967, 1968a, 1968b, 1969; Votaw, 1936); and their general levels of confidence in the testing situation.

6. Glass and Wiley (1964) have presented both theoretical and

empirical evidence in support of the position that Rights-scored tests are more reliable than Formula-scored tests. On the other hand, Lord (1963, 1975) and Lord and Novick (1968, p. 308) argue effectively for the position that Rights-scored tests are less reliable than Formula-scored tests.

7. Arguments and data have been introduced by several investigators--e.g., Cross and Frary, 1977; Cureton, 1966; Rowley and Traub, 1977; and Slakter, 1968a, 1968b--in support of the assertion that the Formula-score directions and advice given in connection with Formula-scored tests is bad advice; examinees would be better advised to guess, even in a Formula-scored test, since their scores would be higher than if they did not guess. This would suggest that in general students know more about the content of a test than they think they do. An implication of this position is that students tested with Formula directions are put at a disadvantage relative to those tested with Rights directions. (This assertion will be referred to later in this paper as the Differential Effect Hypothesis, since its claim is that Formula directions tend to reduce some examinees' scores artifically.)

8. If the assertion in (7) is supported by data, and if it is also true that low-scoring examinees are less inclined to guess than are higher-scoring examinees, then it would follow that Formula directions and Formula-scoring tend to depress low scores still further. It would therefore also follow that Formula directions and Formula-scoring would be especially disadvantageous to minority students, who earn lower scores, on the average, on most tests. Ebel (1968) found in fact that low-scoring examinees are slightly more inclined to guess than higher-scoring examinees. But additional confirmatory data would be useful.

9. The Differential Effect Hypothesis also holds that students who are less inclined to take a risk, and therefore less inclined to guess at items they are not sure of, are further disadvantaged by the Formula-directed test (Sherriffs and Boomer, 1954; Slakter, 1967, 1968a, 1968b, 1969; Votaw, 1936).

10. The argument has been advanced that Rights scores provide discriminations below mean chance, a point which is defined as a zero raw score, in Formula-score terms. Boldt (1968) and Levine and Lord (1959) have in fact demonstrated that there is a valid discrimination provided by scores below mean chance, although not as much as is provided elsewhere on the score scale. In recognition of this fact, some testing programs report scaled scores corresponding to negative Formula raw scores--i.e., Formula scores below mean chance. The perhaps obvious point should be made here that "mean chance" is the mean raw score one would earn if all items one answered were answered at random. However, this is not to say that scores at and below "mean chance" were in fact earned by responding at random.

On the other hand:

1. There have been objections to Rights (permissive) directions on the ground that they encourage indiscriminate guessing, especially when the examinee has insufficient time near the end of the test to consider the remaining items carefully. The argument goes on to say that indiscriminate guessing is educationally deplorable, because it focuses the student's interest on improving his or her test score, without sufficient regard to the educational outcome being assessed by the test. That indiscriminate guessing occurs, especially toward the end of the test, cannot be denied.

Numerous instances have been observed of "pattern-marking" the last several items in tests scored Rights, indicating clearly that random responding has taken place.

2. Lord (1963) has advanced theoretical arguments that because of the random guessing component, which adds error variance of its own, Rights-scored tests are less valid than Formula-scored tests; and under the assumption that omitted items would be replaced with random responses, Rights-scored tests are also less efficient than Formula-scored tests (1974). Lord also demonstrates (1977) that difficult tests are extremely unreliable for low-scoring examinees because they guess more often than they should and, according to Ebel (1968), more often than do other examinees. These tests, Lord finds, would be more reliable for low-scoring examinees if the tests were shortened by removing the more difficult items.

3. Further, there is some concern that data provided by items placed near the end of the test will yield poor estimates of ability in a permissive (Rights-directed) situation because some relatively able examinees will mark the last few items (those that they do not have time to consider more carefully) at random, thereby causing those items to appear to be more difficult for examinees at that level of ability--and more error-laden--than they actually are.

4. With regard to the assertion that examinees are better advised to guess than to omit, on the basis that they generally know more than they think they do, there is evidence to show that this is a more complex issue than it may appear to be. While it is entirely possible that some examinees, those who are less prone to take risks but who have partial information, may

improve their Formula scores by guessing, there are others, especially lower-scoring students, who would do worse, because their information is often misinformation. That this is true is shown by the preponderant numbers of below-chance asymptotes sometimes observed in item response curves (Lord, 1980, p. 110). It is also observed in the large numbers of items with smaller-than-chance proportions of correct responses made by low-ability examinees, as evidenced in typical item analysis outputs. Finally, the data, referenced above, by Boldt (1968) and by Levine and Lord (1969), demonstrate that below-chance scores are not merely random departures from a chance mean; they are valid scores, earned, very likely, as a result of misinformation. Students scoring at those levels would have profited, even on a Formula-scored test, by guessing blindly or by omitting items that were too difficult for them rather than by attempting to answer them.

5. The advantages to the examinee of Rights directions are not as clear as has been claimed. Even when instructed to guess when they are in doubt, some examinees will fail to respond to every item. Whether these examinees leave some items blank because they do not understand the instructions, because they do not trust them and perceive a risk in responding, or because they do not have enough time to respond is not known. The fact remains, however, that Rights directions do not insure that examinees will respond to every item.

6. It has been suggested (L. R Tucker, personal communication) that Formula-scoring as such may compensate for differences in instructions, that even when different students (of equal ability) have been given

different instructions regarding guessing, Formula-scoring will tend to equalize the scores. This is a major consideration. Since the directions given to examinees are intended to influence guessing strategies, this hypothesis further suggests that Formula-scoring may also tend to compensate for individual differences among examinees with respect to their guessing strategies. If this is so, it would argue that tests should be Formula-scored; and consistent with that method of scoring, tests should be administered under restrictive instructions with respect to guessing. (This hypothesis will be referred to later in this paper as the Invariance Hypothesis, since its claim is that Formula scores are invariant with respect to directions for guessing.)

In recent months serious consideration has been given to the pros and cons outlined above in evaluating the desirability of moving from the Formula-directions-and-Formula-scoring mode, which characterizes most of the large-scale testing programs administered by Educational Testing Service today, to the Rights-directions-and-Rights-scoring mode. An important factor in contemplating such a change, however desirable such a change may be from the other points of view, is whether it can be effected without a discontinuity in the score scale. For example, while it certainly is true (as was discussed above) that for a given set of individuals to whom a test has been given under one of the two types of directions, the correlation between the two methods of scoring will be (spuriously) very high, it is also true that the mean and standard deviation of these scores will be quite different: Formula-scoring will inevitably show a lower mean and higher standard deviation of raw scores than Rights scoring. However, a good deal more than a simple shift in mean and standard deviation is involved here. Since the mode of scoring cannot be arbitrarily chosen but must be consistent with the particular types of directions used in the test administration itself, we must necessarily deal with the issue as one involving a conversion of a test administered with Rights directions and scored Rights to the scale of a test administered with Formula directions and scored Formula. Not only the scoring method but the strategic orientation of the examinee is at issue; and it is the combined effects of the behavioral and scoring components of the change that cause a shift in the scores and a possible discontinuity in the scale. Such a discontinuity, if it occurred, would mean that scores earned after the shift would not be strictly comparable to those earned before the shift.

The present study was undertaken, therefore, in an attempt to investigate the psychometric feasibility of changing from the present restrictive/Formula mode to the permissive/Rights mode without endangering scale continuity. In the process of planning the study, care has been taken in the design of special administrations, and analyses based on data resulting from those administrations, to investigate various methods of equating that could be realistically undertaken in the context of an operational testing program, but also to collect data which are designed to cast some light on one or more of the issues on which Formula-score and Rights-score adherents are presently divided.

Technical Problems in Equating Rights Scores to Formula Scores

It is generally known that all of the large-scale testing programs administered by Educational Testing Service report scores to examinees and to test users on a continuing equated scale, a scale which, in nearly all instances, is clearly different and distinguishable from the raw scores of any particular test form. The purposes behind the development and maintenance of this scale are clear. The most important of these purposes is that the scores of students taking the tests at different times and at different administrations are thus made directly comparable, in spite of the inevitable variations in difficulty that exist from one form to another; and no examinee is put at an advantage or disadvantage in relation to other examinees because of the particular level of difficulty of the test form that he or she happens to take. From the score user's point of view this system also has distinct and important advantages: It offers the freedom of using the scores from whatever test form may be conveniently available and free of compromise. There is, additionally, the advantage that scores can be compared across students and groups of students without regard to the test forms that yielded those scores; and special studies can be carried out, aggregating data across forms, plotting trends, and studying the effects of intervening time and treatments on scores. These freedoms are all made possible by a complex system of form-to-form equating to the underlying scale ordinarily carried out immediately following the administration of each new form of the test in each of the large-scale testing programs.

Although there are, as expected, variations and modifications in particular details, one can categorize the equating methods used in the large-scale programs as falling into one or the other of two types. In one of these, the new form is "spiralled"[*] with one or more old forms, and administered, one form to each examinee taking the test at the administration in which the new form is first offered. Using statistics based on the groups of examinees taking the two or more forms, the scores on the new form are equated directly to the scores on the old form(s) in a procedure based on the assumption that the separate groups are equivalent in all respects, but in particular, with respect to the distribution of the abilities in question.

The foregoing method of equating is feasible, however, only in those programs in which more than one test form may be administered at one time. In other testing programs it is customary for all examinees taking the test at a given administration to take precisely the same operational test form. As a result--and since examinees taking the test at different administrations are known to differ--it is not possible to identify equivalent groups of students to use for equating. As an alternative device, the method of equating used in these circumstances employs a set of "common items." These are items that were administered to the examinees when the old form was first given and are given again to other examinees at the time they take the new form. The sense of the "commonality" of these items goes beyond the printed test, however; it also implies, and requires, that the conditions of measurement--the psychological task represented by the items--be the same for both groups of students, because it is on the basis of their observed performance on these items that statistical adjustments are made to compensate for the fact that the groups may not be equivalent. It is the conditions of measurement, clearly, in addition to the content of the items themselves, that also account for performance. The intent of this method of equating is to simulate by statistical adjustments the random (actually, systematic) sampling method described above.

---

[*] "Spiralling" is a term employed at ETS to describe a method of distributing test forms to obtain systematic samples of examinees. When there are two forms to be administered, the forms (say Forms X and Y) are packaged in alternating sequence--X, Y, X, Y, X, ... --and distributed to successive students as they are removed from the top of the package of test books. When there are three forms (X, Y, and Z), they are packaged X, Y, Z, X, Y, Z, X, Y, ..., and similarly distributed. When the total group of individuals to be tested is of size N, and there are m tests, or test forms, to be spiralled, there will be N/m complete spirals, or cycles, of test forms to be distributed, and the *m*th individual in every complete cycle will receive the same test form. Thus, if, for example, seven forms are spiralled, the 3rd, 10th, 17th, 24th, 31st, ... individuals in the group will receive the same form of the test. When the test books are separated by form, the samples of individuals, each receiving a particular test form are (except in highly unusual circumstances) essentially stratified samples and more nearly equivalent than if random sampling methods were employed.

Both of these general methods of equating can operate, and have operated, quite well in the context of secure testing programs, in which it has been possible to protect old forms from compromise so that they can be used again without giving the new groups of students special advantages. In the context of an open-disclosure environment, as has been enacted in New York State, however, severe constraints are imposed on the methods of equating. The first method of equating, which depends on the spiralled administration of the new form along with one or more old forms, is no longer possible, since the old forms would not be secure. The alternative method, which involves a set of "common items," is possible only when the "common items" are nonoperational, that is, do not count toward the examinee's score. That method, it should be noted, is feasible only within the latitude permitted by the present New York State law, since nonoperational items are protected from disclosure by the New York law; laws presently being considered in other jurisdictions may not permit this latitude.

The "common item" method has worked quite satisfactorily in the past when the content of the test and the conditions of administration can be adequately represented in miniature in the set of common items. But when, as would normally be the case in considering a shift from Formula to Rights scoring, the old form and the common items are administered with restrictive instructions and scored by Formula, and the new form and the common items are administered with permissive instructions and scored Rights, and, finally, when the groups taking the forms in these two administrative modes are not in any sense randomly equivalent groups, the usual methods for equating are inapplicable. The scores on the common items do not have the same meaning in the two contexts. What remains is the possibility that the Invariance Hypothesis, defined above, can be utilized in the equating process. This hypothesis, it is recalled, states that differences in the scores earned by two randomly selected groups who have been given different instructions to guess tend to be minimal when the test papers for the two groups are scored by Formula. The principal purpose of the present study is to test this hypothesis. If the data support it, then the sets of common items for the examinees taking the new form--those receiving Rights directions--can be rescored by Formula, allowing direct comparisons between the two groups to be made on the common items in Formula-score terms in the process of equating the new and old forms. Even if the hypothesis is not fully supported, it is possible that the data of the study will provide information to aid in developing appropriate adjustments to overcome the remaining bias.

Still another possibility remains open, although it too represents some risks. If it can be shown that examinees can shift, when they are instructed to do so, from one set of test-taking strategies to another, then the new forms to be administered in the future may be administered under Rights directions and scored Rights, and the sets of common items (if they appear as a separately-timed block) administered under Formula directions and scored by Formula. Such an administration would enable the direct comparison of Formula scores on the equating sections. The risk here is that examinees may be able to identify the nonoperational section because of its different instructions, and perform on it at a reduced level of motivation, attention, and care.

An additional concern, alluded to earlier, in the equating of Rights-administered-and-scored tests to Formula-administered-and-scored tests is the possibility of introducing an additional dimension into the measurement, which may effectively result in the "equating" of nonparallel tests. That is, if the data suggest that the tests represented substantially different psychological tasks, then there will be some considerable question as to the generality of the meaning of the "equating," however it is carried out.

A STUDY OF COLLEGE BOARD SAT-VERBAL AND CHEMISTRY TESTS, BASED ON SPECIAL TEST ADMINISTRATIONS

Questions Addressed by the Study

The principal purpose for which this study was undertaken was to investigate the effectiveness of several methods of equating scores that had been earned under conditions of Rights directions and scoring to scores earned under conditions of Formula directions and scoring. Information on this subject is of vital importance if an operational testing program that has administered and scored its tests in the Formula mode is to be capable of shifting to Rights directions and scoring without introducing a discontinuity of its score scales.

In the course of studying the equating methods it was deemed necessary to investigate other related questions:

1. To what extent do the results provide a firm basis for choosing between the Invariance Hypothesis and the Differential Effect Hypothesis?

The Invariance Hypothesis and the Differential Effect Hypothesis differ essentially in their predictions regarding how well students would perform if, instead of choosing to omit certain items when tested under Formula directions, they chose to answer them. The Invariance Hypothesis implies that their performance on the omitted items would be, on the average, neither better nor worse than would be expected by chance. The Differential Effect Hypothesis, on the other hand, implies that their performance on those items would be better, on the average, than would be expected by chance. If the Invariance Hypothesis is true, Formula scores would remain the same, on the average, whether or not the students chose to omit items about which they had insufficient basis for answering. If, however, the Differential Effect Hypothesis is true, students who choose to omit certain items when tested under Formula directions would be at a disadvantage in comparison with other students of equal ability who answered all the items.

Although the same student cannot take the same test under both Rights and Formula directions at the same time, it is possible to administer the same test so that one random half of a large group is tested with Rights directions and the other half is tested with Formula directions. The Invariance Hypothesis would predict that the two groups would have virtually equal mean Formula scores; the Differential Effect Hypothesis would predict that the group tested under Rights directions would have a higher mean Formula score than the group tested under Formula directions.

2. To what extent do Formula directions affect the number of items considered but intentionally omitted, the number of items not reached, and the total number of items not attempted?

3. To what extent do students comply with the instructions given to them and change their strategies with respect to guessing consistent with those instructions?

4. When students are stratified on the basis of ability, is there a discernible difference between high- and low-ability students in the effect of Formula and Rights directions on the average number of items omitted, not reached, or not attempted? Do Black students show the same results as the total group?

5. Does a guessing index defined as "Wrongs minus Omits" provide useful information about guessing tendencies that is not provided by the various indices of nonresponse?

6. To what extent do Formula and Rights directions yield different reliabilities, as determined by internal consistency and parallel-form methods?

7. Is there reason to believe that the assumption of parallelism between a test administered with Rights directions and the same test administered with Formula directions is not warranted?

8. How much confidence can be placed in the Invariance Hypothesis as a basis for equating Rights scores to Formula scores? To what extent does the use of the Invariance Hypothesis result in systematic differences between conversion lines obtained by assuming invariance and corresponding parameters obtained by traditional equating methods?

Study Design

Several of the previous studies of Rights and Formula scoring--for example, Cross and Frary, 1977; Sherriffs and Boomer, 1954; Slakter, 1968; and Votaw, 1936--have called for the administration of a test under Formula-score directions followed either immediately, or after some intervening time, by a redistribution of the original answer sheets with instructions to review all previously unanswered items and to fill them in, using a differently colored pencil, with a considered or guessed response (Rights-score directions). In no study that we know of was the order of administration counterbalanced, to determine whether these instructions were subject to an order effect. In all these studies the students were, obviously, given additional time to reconsider their previously omitted responses, more time, in aggregate, than a normally administered test with Formula-score directions would have called for. This is a condition of the studies that, by itself, would only have had an artificially elevating effect on their scores. And, finally, as Lord (1975) points out, the Slakter (1968a) study in particular is flawed because the students were allowed several days before the redistribution of answer sheets, during which time they were at liberty to compare notes with one another or to check on doubtful items.

Unlike the foregoing studies, the present study was designed to achieve symmetry, but did not require the examinees to review and respond to a test they had previously taken. It is indeed the only study we know of in which the same test was administered under both Rights and Formula directions. Moreover, the administration of the tests was so arranged that all comparisons would be made between and among experimentally equivalent groups. Most of the analysis was based on special administrations of a form of the College Board SAT-verbal Test that had first been introduced in April 1976, to be referred to in this paper as Form A. Like other current forms of the SAT, Form A is administered in two separately-timed half-hour sections, 45 items in the first section and 40 items in the second section. Both sections contain items of four types: antonyms, analogies, sentence completion, and reading comprehension. Current operational practice is to administer the SAT-verbal with restrictive instructions and to score it by Formula, $R - \frac{1}{4}W$, since all items are five-choice. Four sets of directions for administration were prepared for the present study, identical in all respects except for instructions regarding guessing. The following table describes how the four sets of instructions for Form A were administered. Again, it is recalled that Rights directions are permissive with respect to guessing; Formula directions are restrictive with respect to guessing.

Directions for Administering SAT-verbal, Form A

Set	Part (Section) 1	Part (Section) 2
1	Rights	Rights
2	Rights	Formula
3	Formula	Rights
4	Formula	Formula

Additional, confirmatory analyses were based on the administration of a second form of the SAT-verbal, Form B, a form which was first introduced operationally in June 1976. As just suggested, the analyses based on the administration of Form B were not intended to be as detailed as those based on Form A. They were undertaken to provide assurance that the results of the main analyses were not idiosyncratic to Form A. Form B was constructed to parallel Form A in content, item type, number of items, difficulty, discrimination, and speededness. Two sets of instructions were prepared for the administration of Form B, identical in all respects to those in Set 1 and Set 4, above. The following table describes the instructions for the administration of Form B.

Directions for Administering SAT-verbal, Form B

Set	Part (Section) 1	Part (Section) 2
5	Rights	Rights
6	Formula	Formula

Special test booklets were prepared for each of the six sets described above. Four of the sets of test books contained Form A items, the other two contained Form B items. Each set contained Rights and Formula instructions for Part 1 and/or Part 2, as described above. Inasmuch as the timing for all of the six sets was identical, it was possible to administer all six to different students in the same testing room at the same time. It also permitted "spiralling" the test books in the order: 1, 2, 3, 4, 5, 6, 1, 2, 3, 4, ..., and the distribution of the test books to the students in that order, with the result that every sixth student received the same test book. By means of this procedure of (systematic) sampling, it was possible to achieve very nearly equivalent groups of examinees taking the different test-and-instructions. In fact, the groups formed with this method of sampling were more nearly equivalent than would have been obtained with random sampling methods.

The sample of students chosen for the administration of the SAT-verbal was drawn from a population of students who were likely to be taking the SAT for admission to college. (The specifics of the definition of that population and the selection of schools are described in the following section.)

As a further check on the results of the main analysis, based on Form A of the SAT-verbal, the Chemistry Achievement Test of the College Board Admissions Testing Program was also administered, but to an entirely different sample of students, a sample drawn from students taking first-year chemistry in high schools that have relatively large numbers of students who take the College Board Achievement Tests. The form of the Chemistry Test used in this study was one first introduced in January 1969. Although not as new as the SAT-verbal forms referred to above, this form of the Chemistry Test was re-examined and judged suitable for the experiment as well as for current operational use.

Unlike the SAT, the Chemistry Test is administered under a single one-hour time limit. But for that difference, the test books for Chemistry were prepared in the same way as were the test books for SAT-verbal, Form B. Two types of books were prepared, containing identical items, but differing with respect to instructions, as shown in the table below.

Set	Directions for Administering the Chemistry Test
7	Rights
8	Formula

In those schools chosen, and agreeing, to take the Chemistry Test, test books 7 and 8 were spiralled, so that every odd student took one and every even student took the other.

In preparation for the later anlyses, eight groups of experimental subjects were formed, each corresponding to the eight sets described above, and designated accordingly.

The task of the test supervisor in both types of administrations was limited to the presentation of the following information and instructions: informing the students regarding the fact that they would have different types of instructions to guess; instructing them with regard to the required procedures for identifying themselves and for marking the answer spaces; asking for the identification of sex, the identification of ethnic group (American Indian, Asian American, Black/Afro American, Caucasian, Chicano/Mexican American, Puerto Rico/Puerto Rican American, Spanish American, or Other); asking for the students' rank in class (to the nearest fifth); and, finally, timing the test.

The students participating in the study were also asked whether they wanted their scores to be sent to them and their high school. If they did, their names and their scores on the SAT-verbal (or Chemistry) scaled score scale were reported to their high school, with instructions to the high school to transmit the scores to the students.

Sample Characteristics

The sample design called for testing relatively large numbers of students for whom the test would be appropriate and who were willing to participate. It also called for including a relatively large proportion of minority group members, particularly Black students, in the SAT-verbal sample. Because the SAT-verbal sample was divided into six subgroups, a target figure of 9,000 students, including 2,000 minority students, was used in planning the SAT-verbal sample. For Chemistry, the target figure for sample size was 2,000. Because the study was concerned with equating and subgroup comparisons, which do not require a typical cross-section of examinees, the sample design, as described, was considered to be appropriate.

The selection of schools for the SAT-verbal sample took account of relevant data about schools available in the College Board statistical data file. The selection process used a listing of all schools having 50 or more College Board candidates in 1978-79 and having a school code ending with either of two (randomly selected) last digits. For each school having 100 or more candidates, percent minority and percent Black were also listed.

The first step in designing the sample called for estimating the number of prospective examinees in the group of schools to be invited. Because of limitations in available data on schools, because of uncertainty regarding the proportion of schools and examinees that would choose to participate, and because the time schedule was too tight to permit much replacement of schools that decided not to participate, it was decided to print substantially more test books than the minimum provided for in the study design. Thus, if the participation rate turned out to be unexpectedly high, sufficient test books would be available to permit all schools to test. For SAT-verbal, 15,000 test books were ordered and for Chemistry 3,6000 test books were ordered. Planning for SAT-verbal was based on the number of SAT-takers in 1978-79. In selecting the schools to be invited, we felt that we could safely invite schools having 17,000 SAT-takers and still be able to have a 10% overage in shipments to schools. For Chemistry, it was estimated that about one-third of the 11th grade students in a school would be taking that subject. Accordingly, it was decided to invite schools having a total estimated enrollment in the 11th grade of about 10,800.

In designing the SAT-verbal sample, special steps were taken to insure that a sufficient number of minority group members would be included to provide an adequate sample for separate studies. An exploratory survey of the data available on the school lists for SAT-verbal indicated that this objective could be achieved if approximately half of the prospective examinees were enrolled in the schools having the highest percentage of Black students among their College Board test takers. The other half of the group of prospective examinees could then be obtained from the other schools on the lists. Accordingly, the initial sample included 49 schools having 17% or more Black students among their College Board test-takers and 60 schools selected at random from the remaining schools on the lists. On the basis of the 1978-79 school data, it was estimated that about 24% of the prospective examinees in the SAT-verbal sample would be minority group students and about 18% would be Black students.

The remainder of the SAT sample was selected by random sampling from the total list of schools, excluding Alaska, Hawaii, and the 49 schools already selected. There were 635 schools in the eligible group. It was decided to include 60 schools, enrolling an estimated 8,631 SAT-takers, in the list to be invited. Thus, the total SAT sample included 109 schools and approximately 17,000 students.

The Chemistry sample was selected from a list of 61 schools having a school code with the same (randomly selected) last digit and having 25 or more Achievement Test-takers in 1978-79. Enrollment estimates for 11th grade students were available for 58 schools from the current Preliminary Scholastic Aptitude Test data files. It was decided to impose the further requirement that the number of Achievement Test-takers (as seniors) should be at least 15% of the 11th grade enrollment. As it turned out, the 33 schools meeting these requirements had an estimated total 11th grade enrollment of 10,541. As a result, these 33 schools were selected for the Chemistry sample.

On the basis of a preliminary survey of the returns from the initial mailing, it was decided to augment the sample by inviting additional schools to participate. Because of the tight schedule, these schools were selected only from among those located in New Jersey or in nearby states. Of the 50 supplementary schools for the SAT-verbal sample, 15 schools having a percentage of Black students of 7.0 or higher were selected using the lists for both SAT-verbal and Chemistry. The remaining schools were selected at random from the two school lists prepared for the SAT-verbal sampling. The 20 supplementary schools for Chemistry were selected from schools on one of the SAT-verbal lists that had not been selected for that study. The supplementary sample was selected at random from schools having 25 or more Achievement Test-takers.

Of the 109 schools included in the SAT-verbal sample, 52 provided usable data for the study. The supplementary sample provided data for 17 schools. For Chemistry, 19 of the 33 schools included in the initial sample participated, and nine of the schools in the supplementary sample participated.

Description of the Samples

The 69 schools in the SAT-verbal sample are located in 19 states. New York was represented by 14 schools, and California and Pennsylvania were each represented by nine schools. The 28 schools in the Chemistry sample were located in 12 states and the District of Columbia. Six of the 28 schools were in Massachusetts, with five in New Jersey and four in Pennsylvania.

Characteristics of the students included in the SAT-verbal and Chemistry samples are shown in Table 1. Of the 6,260 students in the SAT-verbal sample, 1,172 belonged to the Black/Afro-American group and 257 were members of the three Hispanic subgroups. Although the sample size both for the total group and for minority group members was smaller than had been planned, the groups were sufficiently large to provide a useful data base for the study. A large percentage (58%) of the participants were female students, and 92% reported that they were in the upper three-fifths of their classes academically.

Approximately half (595) of the 1,172 Black students in the total SAT-verbal sample were enrolled in six schools, each of which enrolled 50 or

Table 1

Distributions of Ethnic Group Membership, Sex, and Rank in Class in SAT-verbal and Chemistry Samples

Characteristic	SAT-verbal		Chemistry	
	N	%	N	%
Ethnic Group Membership				
American Indian	35	0.6	9	0.4
Asian American	121	2.0	65	2.9
Black/Afro American	1172	19.2	36	1.6
Caucasian	4291	70.2	2009	90.9
Chicano/Mexican American	141	2.3	4	0.2
Puerto Rican/Puerto Rican American	61	1.0	7	0.3
Spanish American	55	0.9	11	0.5
Other	230	3.8	68	3.1
Missing Data	154	---	97	---
Total	6260	100.0	2306	99.9
Sex				
Female	3619	58.0	1084	47.1
Male	2623	42.0	1216	52.9
Missing Data	18	---	6	---
Total	6260	100.0	2306	100.0
Rank in Class				
High Fifth	2004	32.5	831	36.6
Second Fifth	1939	31.5	810	35.7
Third Fifth	1724	28.0	514	22.7
Fourth Fifth	375	6.1	76	3.4
Low Fifth	118	1.9	37	1.6
Missing Data	100	---	38	---
Total	6260	100.0	2306	100.0

more members of the Black student sample. Another one-third (391) were enrolled in 12 schools that had from 20 to 48 sample members. The remaining 187 students were enrolled in 34 schools. Twenty-seven schools did not test any Black students for the study.

In the Chemistry sample, about 91% of the tested group who reported ethnic group membership were White. About 53% were male, presumably reflecting a greater tendency for males than for females to enroll in chemistry courses. For Chemistry, over 72 percent were in the top two-fifths in self-reported Rank in Class, and only 5% were in the bottom two-fifths.

Test Administration*

In preparing the test booklets, particular attention was given to the directions with respect to guessing. Because spiralling within school was considered essential to making subgroups receiving different directions as comparable as possible, the supervisor's instructions had to be appropriate for both kinds of directions. At the beginning of the SAT-verbal testing, the supervisor read the following statement:

> You are about to take part in an experiment concerned with the College Board Scholastic Aptitude Test being conducted by the Educational Testing Service, the organization that constructs the College Board SAT and Achievement Tests. The experiment, which will be extremely important to students taking the tests in future years, is being done in order to learn more about the effect of test directions on your test performance.

The statement for Chemistry examinees was the same except for the name of the test.

Just before the SAT-verbal examinees began work on the first section, the supervisor read the following statement:

> This test includes two separately-timed one-half hour sections. Each section has special directions concerning guessing. Some of you will have the same directions concerning guessing for both sections; others will have different directions for the two sections. Please read the directions for each section on your test booklet carefully, and answer the questions in each section according to the directions for that section.

At the beginning of the second separately-timed section, the supervisor again instructed the students to read the directions for the section carefully, and allowed time for them to do so. Directions concerning Formula and Rights scoring were printed on a separate page from other directions for the tests in order to emphasize their importance.

---

*In this discussion separately-timed parts of SAT-verbal are referred to as sections.

The corresponding statement for the Chemistry examinees, given by the supervisor as soon as the test booklets were distributed was as follows:

> This is a 90-item, one-hour test. Please read the directions in your test book carefully, and answer the questions according to the directions.

The Chemistry examinees were instructed to read the directions about guessing as soon as they opened their test booklets; the SAT-verbal examinees were instructed to read the directions about guessing just before they began work on each section.

The Rights directions for SAT-verbal were adapted from the Rights directions used in the Law School Admission Test. They were as follows:

> Read the directions below carefully, and answer the questions in this section according to these directions.
>
> Your score on this section will be based on the number of questions you answer correctly. No deduction will be made for wrong answers. You are advised to use your time effectively and to mark the best answer you can to every question, regardless of how sure you are of the answer you mark.

The Formula directions were essentially the directions used for operational administrations of SAT-verbal as follows:

> Read the directions below carefully, and answer the questions in this section according to these directions.
>
> Students often ask whether they should guess when they are uncertain about the answer to a question. Your score on this section will be based on the number of questions you answer correctly minus a fraction of the number you answer incorrectly. Therefore, it is

> improbable that random or haphazard guessing will change your scores significantly. If you have some knowledge of a question, you may be able to eliminate one or more of the answer choices as wrong. It is generally to your advantage to answer such questions even though you must guess which of the remaining choices is correct. Remember, however, not to spend too much time on any one question.
>
> Do not worry if you are unable to finish this section or if there are some questions you cannot answer; many students leave questions unanswered. You should work as rapidly as you can without sacrificing accuracy. Do not waste time puzzling over a question which seems too difficult for you.

The special directions for the Chemistry examinees were precisely the same except that the word "test" was substituted for the word "section."

Supervisor's manuals were prepared for SAT-verbal and for Chemistry. These manuals were adapted from the manuals used with regular College Board Tests. As suggested by the ETS Board of Prior Review, students were informed by the supervisor at the beginning of the testing session that their participation in the testing and in answering the questions was strictly voluntary and that each student's scores would be reported only to his or her school and to the student and would be reported only if the student requested it by marking an appropriate space on the answer sheet. Students were informed that the experiment was being done "in order to learn more about the effect of test directions on your test performance."

All schools were asked to administer the tests between April 15 and April 30, 1980. For SAT-verbal, "juniors who are planning to take the SAT" were defined as the group to be tested; for Chemistry, "students who are currently enrolled in the second semester of a one-year course in Chemistry" were defined as the appropriate group. Within these general guidelines, each school devised its own procedures for inviting appropriate students to participate and for scheduling and administering the tests.

Except for the analyses of equating methods, analyses of Formula score data were carried out using exact--i.e., unrounded--Formula scores. Because the equating portion of the study was developed to guide operational practice, the analyses of equating methods made use of rounded Formula scores, in which exact scores ending in .5 were uniformly rounded upward, as they are in operational practice at ETS.

Results

Effect of Directions on Mean Formula Scores: Method of Analysis

One set of analyses was designed to test the hypothesis that the mean score for students examined under Rights directions will be equal to the mean score for students examined under Formula directions when Formula scoring is used for both groups of students. This hypothesis, it is recalled, is referred to as the Invariance Hypothesis. If, as Slakter and others have maintained, students tend to omit questions about which they have useful partial knowledge, we would expect that the Invariance Hypothesis would be rejected, and that the mean Formula scores for students tested under Rights directions would be significantly higher than the mean Formula scores for students tested under Formula directions.

The study design called for giving one group of students a test with Formula directions and giving a comparable group of students the same test with Rights directions. The experiment was designed to make the groups receiving different directions as similar as possible in all other respects. The use of the method of spiralling, described earlier in this report, when applied to large groups, tends to produce groups that are quite similar, not only in the abilities measured by the tests but in other respects as well. Moreover, this method, as used in this study, tends to insure that, within each participating school, the number of students receiving each type of directions will be very nearly equal.

In planning the data analysis, it was recognized that the use of a simple t-test of means and would not take account of the fact that examinees were assigned to samples by spiralling rather than by simple random sampling. Under the conditions of this study, spiralling insured that each participating school would be represented by approximately the same number of students in each sample. Because schools differ appreciably in the ability level of their students, the spiralling method would be expected to yield a smaller variability across samples than would simple random sampling. Thus, a simple t-test would be based on an underestimate of the precision of the experiment. Accordingly, it was decided to use a regression approach, using School Attended to create a set of dummy variables. It was further decided to use Sex, Ethnic Group Membership, and Rank in Class as covariates in analyses of SAT scores in order to increase further the precision of the group comparisons. For this purpose, Rank in Class was analyzed by calculating regression weights not only for the observed values but also for the second, third, and fourth powers of the ranks. Orthogonal polynomials were used in determining the weights for the higher order variables because it was expected that the intercorrelations of the four variables created by computing successive powers would be high. Ethnic Group Membership was so coded as to provide the following categories: Black-White; Hispanic-White; Other-White. Students who did not report Sex, Rank in Class, or Ethnic Group Membership were not included in the analysis of means.

The foregoing analysis plan was applied to Form A and Form B of SAT-verbal separately for Part 1, Part 2, and Total scores, yielding six analyses. In addition, the study design made it possible to perform two analyses using Part 2 scores on Form A as the dependent variable with Part 1 scores as an additional covariate. Finally, one analysis of mean scores was made for the students taking the College Board Chemistry Test. In this analysis, only School, Sex, and Rank in Class were used as covariates, because less than 10% of the Chemistry sample were members of ethnic groups other than White, and because 97 of the 2,306 members of the Chemistry sample did not report ethnic group membership.

Effect of Directions on Mean Formula Scores: Findings

A useful preliminary survey of the results of this study can be obtained by considering the means and standard deviations of Rights and Formula scores for part and total scores shown in Table 2. Within each of the seven sets of results, the spiralling design yields mean scores that may be compared directly with each other.

Considering first the results when Rights scoring is used, there is a consistent pattern for Rights directions to yield higher mean scores than do Formula directions. This result is certainly to be expected on logical grounds because under Rights directions the student's optimal strategy is to answer every item. However, the effect of the directions on the mean Rights scores is relatively small. Only for Total scores on the tests does the difference in Rights scores exceed one raw score point. It is also noted that Rights scores obtained when Formula directions were used tend to have somewhat smaller standard deviations than Rights scores obtained when Rights directions were used.

Table 2

Descriptive Statistics on Rights and Exact (Unrounded) Formula Scores for Part and Total Scores on SAT-verbal and for Chemistry Total Scores for Each Subgroup

							Rights Score[a]		Formula Score[a]	
Test	Form	Part	Number of Items	Group	Directions	N	Mean	S.D.	Mean	S.D.
SAT-V	A	1	45	1	R	1026	*22.22*	*6.93*	17.11	8.30
"	"	"	"	2	R	1068	*22.14*	*6.71*	17.00	8.09
"	"	"	"	1+2	R	2094	*22.18*	*6.82*	17.06	8.19
"	"	"	"	3	F	1054	21.80	6.85	*17.22*	*8.19*
"	"	"	"	4	F	1038	21.50	6.51	*16.86*	*7.78*
"	"	"	"	3+4	F	2092	21.65	6.69	*17.04*	*7.99*
"	B	"	"	5	R	1040	*21.02*	*7.09*	15.66	8.45
"	"	"	"	6	F	1034	20.37	6.94	*15.49*	*8.35*
SAT-V	A	2	40	1	R	1026	*18.24*	*6.42*	13.18	7.81
"	"	"	"	3	R	1054	*18.23*	*6.33*	13.30	7.60
"	"	"	"	1+3	R	2080	*18.24*	*6.37*	13.25	7.70
"	"	"	"	2	F	1068	17.45	6.08	*12.88*	*7.39*
"	"	"	"	4	F	1038	17.48	5.97	*12.84*	*7.26*
"	"	"	"	2+4	F	2106	17.46	6.03	*12.86*	*7.32*
"	B	"	"	5	R	1040	*19.28*	*6.62*	14.42	8.14
"	"	"	"	6	F	1034	18.67	6.52	*14.26*	*8.01*
SAT-V	A	Total	85	1	R	1026	*40.47*	*12.75*	30.30	15.36
"	"	"	"	4	F	1038	38.99	11.90	*29.70*	*14.36*
"	B	"	"	5	R	1040	*40.30*	*13.07*	30.08	15.82
"	"	"	"	6	F	1034	39.03	12.86	*29.76*	*15.68*
Chem.	-	Total	90	7	R	1151	*34.50*	*10.74*	22.06	13.03
"	-	"	"	8	F	1155	32.21	10.26	*21.52*	*11.90*

[a] Means and standard deviations for which the scoring is consistent with the directions appear in italics.

To some extent, the fact that the data were collected under experimental conditions rather than operational conditions may have resulted in smaller effects for differences in directions. In particular, it seems plausible that students would have less incentive to make random responses when the tests are given under Rights directions under the experimental conditions than under conditions of actual (i.e., formal, operational) testing.

The results obtained when the tests were scored by Formula provide a useful preliminary indication of the extent to which the use of Formula scoring removes the effect of different directions. This preliminary analysis suggests that the use of Formula scores reduces but does not eliminate the effect of different directions on the mean scores. With respect to the standard deviation of scores, the use of Formula scoring does not seem to have any consistent effect on the relative size of the standard deviations. Formula directions tend to yield slightly smaller standard deviations than Rights directions for Formula scoring, just as they did for Rights scoring. The problem of standard deviations will be considered again, when equating methods are applied to the data.

Application of regression methods to the data obtained using Formula scoring provides more precise estimates of the effects of differences in directions on mean Formula scores and also provides significance tests for the observed differences. Results for nine analyses based on all members of the designated groups who had complete data on the covariates are presented in Table 3. Of the eight analyses performed

Table 3

Effect of Differences in Directions on Formula Scores: All Students[a]

Dependent Variable							t-test Results		
Test	Form	Part(s)	Number of Items	Covariates	Group(s) Tested Using Rights Directions	Group(s) Tested Using Formula Directions	N	Adjusted Difference[b]	t[c]
SAT-V	A	1	45	School, Sex, Rank, Ethnic Group	1+2	3+4	4013	+0.03	0.17
"	B	1	45	"	5	6	1986	+0.26	0.90
"	A	2	40	"	1+3	2+4	4013	+0.49	2.63**
"	B	2	40	"	5	6	1986	+0.24	0.88
"	A	1+2	85	"	1	4	1985	+0.82	1.64
"	B	1+2	85	"	5	6	1986	+0.50	0.97
"	A	2	40	Sch, Sex, Rank, EthnicGp, Part 1	1	2	2008	+0.29	1.49
"	A	2	40	"	3	4	2005	+0.31	1.71
Chem.	-	-	90	School, Sex, Rank	7	8	2262	-0.00	-.00

[a] Students with missing data on any covariate are excluded.

[b] In all analyses, a plus sign indicates that students tested using Rights directions earned higher mean adjusted scores. A minus sign indicates that students tested using Formula directions earned higher mean adjusted scores.

[c] Significance level, using two-tailed tests:
** $p < .01$

for SAT-verbal, all show that students tested using Rights directions have a somewhat higher mean Formula score than students tested using Formula directions. However, only one of the eight differences attains statistical significance. In the two analyses using Part 1 scores as an additional covariate, the adjusted differences on the 40-item Part 2 were approximately three-tenths of a raw score point. The two results for Total Formula scores on the 85-item SAT-verbal have an average value of about two-thirds of a raw score point. A difference of two-thirds of a raw score point on Form A of SAT-verbal would be equivalent to about 5 scaled score points (since the slope parameter for this form of SAT-verbal is 7.2588), an amount which is more likely than not to be inflated simply because of the greater speededness, and consequent lower mean scores, in a Formula-directed test. But even if this finding were taken at face value in support of the Differential Effect Hypothesis, it must be emphasized that the significance results support the Invariance Hypothesis. From a practical standpoint, the results suggest that assuming the Invariance Hypothesis for equating Rights scores to Formula scores would be unlikely to result in a serious discontinuity in the scale, at least, in the vicinity of the mean.

Results for the single analysis of Chemistry Test data show a difference in adjusted means almost exactly equal to zero. Perhaps the most tenable interpretation of this result is that it is essentially similar to the results for SAT-verbal. The specific outcome for Chemistry suggests that, if anything, the Invariance Hypothesis is more appropriate for subject-matter tests than for aptitude measures.

The data used in the analyses in which Part 1 scores were used as a covariate may also be analyzed using Part 1 as a stratification variable, and assuming that the outcome is not affected by the fact that one of the two groups in each comparison was given different directions for the two sections and the other was not. When stratification is based on Part 1 scores, the results throw light on the question of whether the effect of differences in scoring directions is related to a student's ability level. It is also possible to divide the sample into two groups on the basis of Items Not Attempted in Part 1 administered under Formula directions. This analysis provides a comparison of results for students who chose not to answer a substantial number of items (9 or more) with results for students who chose not to answer relatively few items (8 or fewer). As shown in Table 4, there is no apparent trend in the size of the effect when students are stratified on Rights scores, although students in the highest stratum show the largest effect. When stratification is on Formula scores, there is a trend for effects to be larger for students in the lower strata. When the two results are considered together, they suggest that there is no consistent trend for ability level to be related to the size of the effect of different directions. Results for the groups stratified on the basis of items Not Attempted, when the test used for stratification was administered under Formula directions, are quite similar for students having 9 or more Nonattempts and for those having fewer than 9 Nonattempts. This result would be

Table 4

Effect of Differences in Directions on Formula Scores Earned on Part 2 of SAT-verbal for Groups Stratified on the Basis of Rights Scores, Formula Scores, and Nonattempts on Part 1[a]

Stratification Variable	Interval	Group Tested Using Rights Directions	Group Tested Using Formula Directions	t-test Results		
				N	Adjusted Difference[b]	t[c]
Rights Score	25 and higher	1	2	734	+0.75	+1.69
"	20-24	"	"	600	+0.00	+0.00
"	15-19	"	"	412	+0.42	+0.91
"	14 and lower	"	"	262	+0.34	+0.63
"	Total	"	"	2008	+0.38	+1.41
Formula score	21 and higher	3	4	665	+0.27	+0.63
"	15-20	"	"	613	+0.29	+0.82
"	9-14	"	"	447	+0.37	+0.89
"	8 and lower	"	"	280	+0.72	+1.55
"	Total	"	"	2005	+0.60	+2.36*
Items Not Attempted	9 and higher	3	4	475	+0.67	+1.32
"	8 and lower	"	"	1530	+0.59	+2.00*
"	Total	"	"	2005	+0.60	+2.36*

---

[a] Covariates used in all analyses were: School, Sex, Rank in Class, and Ethnic Group. Students with missing data on any covariate were excluded.

[b] In all analyses, a plus sign indicates that students tested using Rights directions earned higher mean adjusted scores. A minus sign indicates that students tested using Formula directions earned higher mean adjusted scores.

[c] Significance levels, using two-tailed tests:

* $p < .05$

plausible if the Invariance Hypothesis is warranted.

In addition to the analyses based on the total group, two regression studies of SAT-verbal were performed using only Black students (Table 5). In these analyses, it was possible to combine data for two groups who had Rights directions on Form A and for two groups who had Formula directions on Form A. In both analyses, the differences were small and were not statistically significant. As it happened, Black students earned slightly higher Formula scores when tested using Formula directions than when tested using Rights directions. Although these results are opposite in direction to those found for the corresponding total groups, it seems probable that the difference is attributable to sampling fluctuations and that attempts to interpret this difference would not be warranted.

The study design provided that two of the groups taking SAT-verbal would have Rights directions on one part and Formula directions on the other part. Thus, it was possible to compare the performance on Part 2 of students who had different directions on the two parts with the performance of students who had the same directions on both parts. As shown in Table 6, results based on all students show that performance on Part 2 for groups that had different directions on the two parts of the test is remarkably similar to the results for groups that had the same directions on the two parts of the test, indicating that students at this level are quite capable of changing their guessing strategies in the middle of an administration in accordance with changes in directions to guess, as was assumed in the discussion of Table 4. It was

Table 5

Effect of Differences in Directions on Formula Scores: Black Students[a]

Dependent Variable							t-test Results		
Test	Form	Part(s)	Items	Covariates	Group(s) Tested Using Rights Directions	Group(s) Tested Using Formula Directions	N	Adjusted Difference[b]	t
SAT-V	A	1	45	School, Sex, Rank	1+2	3+4	760	-0.23	-0.52
"	A	2	40	"	1+3	2+4	760	-0.10	-0.27

---

[a] Students with missing data on any covariate were excluded.

[b] In all analyses, a plus sign indicates that students tested using Rights directions earned higher mean adjusted scores. A minus sign indicates that students tested using Formula directions earned higher mean adjusted scores.

Table 6

Effect of Directions on Part 1 on Scores Earned on Part 2

Dependent Variable						t-test Results		
Part	Directions	Score	Covariates	Group Tested Using Rights Directions on Part 1	Group Tested Using Formula Directions on Part 1	No. of Cases	Adjusted Difference[a]	t
			Results Based on All Students in Designated Groups[b]					
2	Rights	Formula	School, Sex, Rank, Ethnic Group	1	3	1992	-0.13	-0.50
2	Formula	Formula	"	2	4	2021	-0.11	-0.42
			Results Based on Black Students in Designated Groups[b]					
2	Rights	Formula	School, Sex	1	3	389	-0.97	-1.82
2	Formula	Formula	"	2	4	371	-0.23	-0.42

[a] In this table, a minus sign indicates that students who had the same directions on both parts of the test earned lower mean adjusted scores on Part 2 than did those students who had different directions on the two parts of the test.

[b] Students with missing data on any covariate are excluded.

found that Black students who were asked to shift from Formula directions on Part 1 to Rights directions on Part 2 earned somewhat higher Part 2 scores than Black students who had Rights directions for both parts, although the difference did not attain statistical significance. For the other analysis of shift in directions, the results for Black students were quite similar to those for all students.

Effect of Directions on Nonresponse

Four types of scores, other than Rights and Formula scores, were subjected to separate investigation in this study. These are the number of items Omitted, the number of items Not Reached, the number of items Not Attempted (the sum of the number Omitted and the number Not Reached), and an arbitrarily constituted measure of Guessing, to be discussed in the next section. In defining these variables it is assumed that all items left unanswered prior to the last item reached were in fact considered, but intentionally left unmarked, presumably for reasons of insufficient knowledge, ability, skill, etc. These are the Omitted items. All items left unmarked beyond the last item marked are presumed to be those that the examinee has not had time to consider. These are the items Not Reached.

In order to determine whether there was any relationship between score level and ethnic group and the number of items Omitted, Not Reached, Not Attempted, and Guessed, the tabulations in Tables 7-12 were prepared, one for each of the four groups taking Form A of SAT-verbal, and one for each of the two groups taking Form B of SAT-verbal. Each of these six

Table 7

Number of Items Omitted, Not Reached, Not Attempted, and Guessed on Part 2 of SAT-V, Given with Rights Directions, for White (W), Black (B) and Black plus Hispanic (B+H) Students, Stratified by Rights Scores on Part 1

(Based on Group 1)

Score Interval	Ethnic Group	N	Omitted		Not Reached		Not Attempted[a]		Guessing Index (W-O)	
			Mean	SD	Mean	SD	Mean	SD	Mean	SD
25 and higher	W	316	0.25	1.20	0.34	1.29	0.59	1.78	15.03	5.35
	B	18	0.22	0.55	0.33	1.03	0.56	1.10	17.89	4.59
	B+H	23	0.26	0.54	0.26	0.92	0.52	0.99	18.22	5.13
20 - 24	W	223	0.66	2.10	0.58	1.62	1.25	2.71	20.20	5.68
	B	46	1.07	2.67	0.80	1.98	1.87	3.32	20.78	6.33
	B+H	52	0.96	2.53	0.83	2.02	1.79	3.25	20.69	6.09
15 - 19	W	120	0.80	2.75	1.05	2.19	1.85	3.55	22.22	6.63
	B	61	0.48	2.45	1.61	3.21	2.08	4.59	24.38	6.74
	B+H	79	0.38	2.16	1.34	2.92	1.72	4.13	24.75	6.23
14 and lower	W	35	2.40	5.80	1.37	2.74	3.77	6.44	21.49	11.67
	B	77	1.39	3.46	3.01	4.77	4.40	7.12	24.53	9.41
	B+H	90	1.39	3.39	2.87	4.59	4.26	6.78	24.29	9.16
Total	W	694	0.59	2.29	0.59	1.70	1.18	2.93	18.26	6.84
	B	202	0.94	2.85	1.85	3.69	2.78	5.46	23.04	7.92
	B+H	244	0.86	2.70	1.69	3.51	2.56	5.14	23.10	7.63

[a] Omits plus Not Reached

Table 8

Number of Items Omitted, Not Reached, Not Attempted, and Guessed on Part 2 of SAT-V, Given with Formula Directions, for White (W), Black (B), and Black plus Hispanic (B+H) Students, Stratified by Rights Scores on Part 1

(Based on Group 2)

Score Interval	Ethnic Group	N	Omitted		Not Reached		Not Attempted[a]		Guessing Index (W-O)	
			Mean	SD	Mean	SD	Mean	SD	Mean	SD
25 and higher	W	341	3.32	3.94	0.43	1.34	3.75	4.29	10.26	8.01
	B	26	3.50	4.18	0.69	2.11	4.19	4.72	12.65	8.99
	B+H	30	3.43	4.17	0.77	2.13	4.20	4.96	12.43	8.48
20 - 24	W	233	3.47	4.38	0.94	1.99	4.41	5.19	14.79	8.98
	B	43	2.44	4.50	2.05	2.86	4.49	6.03	16.58	9.10
	B+H	56	3.21	5.20	1.80	2.69	5.02	6.27	15.64	10.30
15 - 19	W	120	3.19	4.71	1.55	2.84	4.74	5.73	17.93	9.97
	B	57	1.86	3.25	1.65	3.22	3.51	5.35	22.25	7.95
	B+H	70	2.03	3.55	1.67	3.14	3.70	5.35	21.56	8.02
14 and lower	W	49	3.45	5.20	2.53	3.38	5.98	6.25	19.45	10.53
	B	65	1.52	3.60	2.40	3.59	3.92	5.44	24.02	8.71
	B+H	78	1.50	3.71	2.18	3.39	3.68	5.38	24.27	8.60
Total	W	743	3.36	4.29	0.91	2.11	4.27	5.00	13.53	9.41
	B	191	2.10	3.83	1.86	3.18	3.96	5.43	20.27	9.50
	B+H	234	2.32	4.18	1.76	3.03	4.07	5.54	19.88	9.81

Table 9

Number of Items Omitted, Not Reached, Not Attempted, and Guessed on Part 2 of SAT-V, Given with Rights Directions, for White (W), Black (B), and Black plus Hispanic (B+H) Students, Stratified by Formula on Part 1

(Based on Group 3)

Score Interval	Ethnic Group	N	Omitted		Not Reached		Not Attempted[a]		Guessing Index (W-O)	
			Mean	SD	Mean	SD	Mean	SD	Mean	SD
21 and higher	W	320	0.54	1.51	0.31	1.17	0.85	2.02	14.51	5.20
	B	20	0.50	0.89	0.55	1.50	1.05	1.93	16.15	3.92
	B+H	27	0.78	1.85	0.59	1.58	1.37	2.42	15.52	3.93
15 - 20	W	217	0.87	2.47	1.04	2.26	1.90	3.54	18.77	5.83
	B	42	1.26	2.55	1.24	2.49	2.50	3.55	19.64	5.50
	B+H	50	1.12	2.39	1.06	2.32	2.18	3.35	19.64	5.29
9 - 14	W	140	1.39	3.15	1.14	2.50	2.54	4.15	21.05	7.09
	B	60	1.78	4.24	1.65	2.59	3.43	5.18	21.78	8.53
	B+H	77	1.84	4.42	1.70	2.75	3.55	5.30	21.53	8.78
8 and lower	W	45	0.91	2.39	0.93	3.03	1.84	4.21	24.38	5.59
	B	77	1.25	3.65	1.99	3.48	3.23	5.78	25.47	8.48
	B+H	90	1.29	3.79	2.31	5.22	3.60	7.00	25.01	9.30
Total	W	722	0.83	2.28	0.73	2.01	1.55	3.23	17.67	6.60
	B	199	1.34	3.47	1.58	2.89	2.92	4.93	22.19	8.14
	B+H	244	1.37	3.60	1.67	3.75	3.05	5.50	21.76	8.50

Omits plus Not Reached

Table 10

Number of Items Omitted, Not Reached, Not Attempted, and Guessed on Part 2 of SAT-V, Given with Formula Directions, for White (W), Black (B), and Black plus Hispanic (B+H) Students, Stratified by Formula Scores on Part 1

(Based on Group 4)

Score Interval	Ethnic Group	N	Omitted		Not Reached		Not Attempted[a]		Guessing Index (W-0)	
			Mean	SD	Mean	SD	Mean	SD	Mean	SD
21 and higher	W	277	2.79	3.32	0.55	1.57	3.34	3.63	10.44	7.09
	B	12	2.33	3.37	0.92	1.73	3.25	4.43	13.33	7.45
	B+H	18	2.06	3.13	1.11	2.08	3.17	4.03	14.33	8.15
15 - 20	W	262	3.61	4.55	1.02	2.18	4.63	5.23	13.93	8.87
	B	45	2.44	3.22	1.18	2.26	3.62	4.09	17.02	6.54
	B+H	58	1.98	2.98	1.09	2.12	3.07	3.81	18.10	6.23
9 - 14	W	128	2.73	3.89	1.80	2.87	4.54	5.31	18.33	8.35
	B	70	2.06	3.56	2.10	4.26	4.16	6.86	21.49	9.16
	B+H	82	1.96	3.44	2.02	4.33	3.99	6.91	21.68	9.09
8 and lower	W	47	1.87	3.27	1.60	2.85	3.47	4.88	23.13	8.20
	B	64	0.83	2.17	1.75	3.39	2.58	4.19	26.30	6.21
	B+H	73	0.86	2.11	1.64	3.24	2.51	4.01	26.23	6.11
Total	W	714	3.02	3.93	1.02	2.21	4.04	4.70	13.97	8.87
	B	191	1.75	3.11	1.69	3.45	3.44	5.33	21.53	8.55
	B+H	231	1.63	2.96	1.60	3.39	3.23	5.20	21.65	8.33

Table 11

Number of Items Omitted, Not Reached, Not Attempted, and Guessed on Part 2 of SAT-V, Given with Rights Directions, for White (W), Black (B), and Black plus Hispanic (B+H) Students, Stratified by Rights Scores on Part 1

(Based on Group 5)

Score Interval	Ethnic Group	N	Omitted		Not Reached		Not Attempted[a]		Guessing Index (W-O)	
			Mean	SD	Mean	SD	Mean	SD	Mean	SD
25 and higher	W	278	0.29	1.10	0.18	0.79	0.47	1.48	13.40	5.19
	B	19	0.79	2.76	0.16	0.69	0.95	2.92	16.74	7.13
	B+H	28	0.86	2.68	0.14	0.59	1.00	2.88	16.39	6.29
20 - 24	W	217	0.41	1.36	0.37	1.19	0.77	1.96	18.61	5.14
	B	28	0.36	0.87	0.54	1.26	0.89	1.73	19.86	4.66
	B+H	34	0.29	0.80	0.44	1.16	0.74	1.60	19.85	4.25
15 - 19	W	146	0.78	1.95	1.09	2.23	1.87	3.24	19.89	6.04
	B	59	0.53	1.79	0.95	2.23	1.47	2.88	23.24	6.72
	B+H	78	0.69	1.92	1.13	2.35	1.82	3.30	22.86	6.69
14 and lower	W	67	1.22	3.46	1.93	3.67	3.15	5.46	22.82	8.82
	B	91	0.47	1.29	1.27	2.78	1.75	2.94	26.90	4.89
	B+H	105	0.50	1.32	1.22	2.66	1.71	2.92	26.70	4.90
Total	W	708	0.51	1.74	0.59	1.80	1.11	2.77	17.23	6.64
	B	197	0.50	1.59	0.96	2.33	1.47	2.78	23.82	6.60
	B+H	245	0.57	1.68	0.96	2.26	1.53	2.91	23.35	6.58

[a] Omits plus Not Reached

Table 12

Number of Items Omitted, Not Reached, Not Attempted, and Guessed on Part 2 of SAT-V, Given with Formula Directions, for White (W), Black (B), and Black plus Hispanic (B+H) Students, Stratified by Formula Scores on Part 1

(Based on Group 6)

Score Interval	Ethnic Group	N	Omitted		Not Reached		Not Attempted[a]		Guessing Index (W-O)	
			Mean	SD	Mean	SD	Mean	SD	Mean	SD
21 and higher	W	245	2.53	2.91	0.56	1.39	3.09	3.26	8.46	6.33
	B	14	2.86	3.11	0.36	0.93	3.21	3.09	9.43	6.24
	B+H	18	2.50	2.87	0.44	1.04	2.94	2.82	10.28	5.81
15 - 20	W	236	3.33	4.01	0.78	1.82	4.11	4.90	12.87	8.28
	B	22	4.14	4.68	0.86	1.61	5.00	4.90	12.82	9.63
	B+H	32	3.91	4.62	0.91	1.75	4.81	4.70	13.09	9.41
9 - 14	W	151	2.74	3.95	1.24	2.38	3.98	5.22	16.64	9.16
	B	56	2.38	4.81	2.05	3.01	4.43	6.36	19.14	10.44
	B+H	72	2.61	4.89	1.85	2.80	4.46	6.26	18.83	10.49
8 and lower	W	78	1.73	3.58	1.21	2.50	2.94	4.96	22.12	9.10
	B	100	1.60	3.12	1.74	2.94	3.34	4.39	24.44	6.65
	B+H	109	1.61	3.24	1.68	2.87	3.28	4.40	24.46	6.91
Total	W	710	2.75	3.63	0.85	1.93	3.60	4.50	13.17	9.07
	B	192	2.21	3.93	1.63	2.77	3.84	5.03	20.47	9.57
	B+H	231	2.31	4.05	1.53	2.64	3.84	5.02	20.03	9.73

[a] Omits plus Not Reached

groups has been stratified on the scores earned on Part 1 of the SAT-verbal, in categories of Rights scores: 25 and higher, 20-24, 15-19, and 14 and lower (and total) for those groups (Groups 1, 2, and 5) for whom Part 1 was administered under Rights directions; and in categories of Formula scores: 21 and higher, 15-20, 9-14, and 8 and lower (and total) for those groups (Groups 3, 4, and 6) for whom Part 1 was administered under Formula directions. The means and standard deviations of items Omitted, Not Reached, Not Attempted, and Guessed were determined from Part 2 data, separately for Whites, Blacks, and Blacks and Hispanics combined.

The means of the Omitted items seem to show uneven, but clearly different trends for the two sets of directions. There is a slightly greater tendency for lower-scoring groups than for higher-scoring groups to omit items in Rights-directed test situations, a characteristic which is shared by all three ethnic groups, and a generally smaller tendency for lower-scoring groups than for higher-scoring groups to omit items in Formula-directed test situations. In the first (Rights-directed) situation, it is probably a fact that lower-scoring groups do not follow the sense of the directions given to them as well as their higher-scoring counterparts do, and guess less often than they should under these directions. In the second (Formula-directed) situation, the same kind of explanation is reasonable: that lower-scoring groups again fail to follow the sense of the directions given to them and guess more often than they should. This tendency may also arise from a greater failure (for them than for higher-scoring groups) to assess their own degree of knowledge and competence on the items. This observation (and speculation) supports the assertions attributed to Ebel (1968) and Lord (1977), supported also by data observed in item analyses, alluded to earlier, that lower-scoring examinees guess more, not less, than higher-scoring students do (Ebel), and more than they should (Lord) on Formula-scored tests.

Considering the data by ethnic group within strata, there appear to be no observable trends to report. However, there are differences in mean Omits and in standard deviations of Omits with respect to the nature of the directions given for guessing. Those groups who were given Formula (restrictive) instructions with respect to guessing in Part 2 of the test (Groups 2, 4, and 6) show much greater average numbers and dispersions of omitted items than those who were given Rights (permissive) directions in Part 2 (Groups 1, 3, and 5).

The means and standard deviations of Items Not Reached do show a clear progression across ability groups, with examinees in the lower ability groups showing greater average numbers of unreached items than those in the higher ability groups. The NR count is often used as an index of speededness and is known to correlate negatively with score. These results, therefore, are not unexpected and lend added support to the reasonableness of the data. What is of special interest here is the fact that the relationships between ability and number of items Not Reached are not spurious here, in the sense that the correlation might be coerced by the constraint of the total number of items. The correlation represented here is that between ability, as measured by the score on Part 1, versus the count of Not Reached independently observed on Part 2.

The same progression of the counts of Not Reached on Part 2, as a function of score level on Part 1, is observed for all three ethnic groups: White, Black, and Black plus Hispanic. There are also some differences (but not quite as clear or sharp, probably because of the small samples of Blacks and Hispanics), among the ethnic groups within strata, with the numbers of Nonreached items largest for Blacks, intermediate for the Black-plus-Hispanic group, and smallest for Whites. These differences are probably attributable in part to differences in score level, even within the strata. Because the strata are relatively broad, it should be noted that even within strata the groups are not precisely matched on ability. As a result it is possible that ability differences within strata may have affected the results.

We also observe that there are small but persistent differences in the numbers of Not Reached items between the tests administered under Rights directions and those administered under Formula directions, with a strong tendency for those tested under Formula directions than for those tested under Rights directions to show a large number of items Not Reached. There are three reasonable, and not incompatible, explanations for this. One is that Formula directions require more time on the part of the examinee; in view of the penalty for incorrect responses, the examinee is obliged to consider and weigh his (her) responses a little more carefully than if the answer sheet were to be scored Rights Only. A second is that students may have engaged in blind guessing under Rights directions near the end of the test, causing an underestimate of the number of items not considered. The third is that the distinction between Omitted and Not Reached items is probably somewhat contaminated, even if not seriously so. The distinction between these two types of Nonattempts assumes that the examinee progresses systematically through the test, responding or omitting as he or she goes along, without skipping or returning to an item considered earlier. Indeed, it is entirely likely that on occasion an examinee will omit an item with the intention of returning to it if time permits--but then time does not always permit. It is also likely that some items classified as Not Reached are in fact considered and intentionally omitted. These explanations may well account for the fact that the data on the Not Reached items show differences between the two types of directions, not as clearly, to be sure, but in the same direction as shown by the counts of Omitted items.

Effect of Directions on Guessing

Considerable thought was given to investigating the extent to which examinees at different score levels guess the answers to the items, and a highly simplified measure of guessing, namely W-O, where W and O are respectively, a count of the number of items answered incorrectly and the number of items omitted, was considered for this purpose. (Note again, that the number of items omitted is taken to include only those items presumed to have been examined, considered, and consciously skipped, apart from those that the examinees are presumed not to have reached within the time limit, those near the end of the test. The items Omitted are the unmarked items followed by one or more marked items; the items Not Reached are the unmarked items that are not followed by any marked items.) The justification for the proposed index, W-O, is that W includes all items for which it may safely be assumed that the student had less than complete knowledge. (W may also include other errors, for example, those arising from failure to understand the directions and from carelessness in marking the answer sheet.) Assuming that the student had no clerical errors in responding, then, it is presumed that any student who responded without complete knowledge did so with at least some degree of guesswork. The subtraction for Omits was introduced as indicating a conscious suppression of any tendency to guess. Because some wrong answers arise from partial or incorrect knowledge, the index cannot be regarded as a pure measure of guessing tendency. However, for students of equal ability, W-O should be a useful indicator of relative tendency to guess.

Other investigators have also developed indices of guessing. A review of these developments and their applications appear in Slakter (1967). Swineford (1938), for example, reports on the development of a measure of "gambling tendency," derived from a special administration of the test, in which she asked the students to rate their level of confidence on a 3-point scale (2, 3, or 4 with the rating of 4 representing high confidence) in responding to true-false items. The formula for her index of gambling is the following:

$$\text{Gambling} = G = \frac{\text{Errors Marked "4"}}{\text{Total Errors} + 1/2 \text{ Omissions}} \times 100.$$

In a later article Swineford (1941) found that the index was "independent of the scores on the tests from which they were computed and also independent of five mental factors [General, Spatial, Verbal, Speed, and Memory] which have been measured by a larger battery of tests." She also found that the "intercorrelations among the [index] scores from [four tests administered to the experimental subjects--Paper Form Board, General Information, Word Meaning, and Deduction--] are sufficiently high to yield a multiple correlation of .85 when all four measures are combined in a regression estimate of the G [guessing] factor" (1941).

Ziller (1957) also offers an index of guessing, but provides no empirical data to test its reasonableness. Unlike Swineford's index, Ziller's is based entirely on the ordinary item responses given by the students in the regular administration of the test, unmodified by special instructions (e.g., expressing degrees of confidence in the responses). Ziller's index is a proportion of the Wrongs to the total of the Wrongs and the items Not Attempted, in the following relationship:

$$Z = \frac{[k/(k-1)]\ W}{[k/(k-1)]\ W + NA},$$

where k = the number of options per item

W = the number of incorrect responses

NA = the number of Nonattempted items, which includes what the present authors refer to as "Omits" plus what they refer to as items "Not Reached."

It was expected, in considering the W-O index, that it would correlate negatively with score level. To determine the strength of this relationship, W-O scores were correlated in Groups 1 and 4 with: a) the scores of the tests from which they were calculated; and (b) the scores of the parallel tests administered in the other half-hour session. Thus, if R, F, and I are taken, respectively, to represent Rights scores, Formula scores, and the index, W-O; and if the subscripts 1 and 2 are taken, respectively, to represent measures derived from the first and second parts of Form A, then the following correlations can be evaluated:

Group 1; N=1026	Group 4; N=1038
$r_{R_1I_1} = -.605$	$r_{F_1I_1} = -.587$
$r_{R_1I_2} = -.533$	$r_{F_1I_2} = -.514$
$r_{R_2I_1} = -.503$	$r_{F_2I_1} = -.496$
$r_{R_2I_2} = -.676$	$r_{F_2I_2} = -.602$

As may be seen in the correlations given above, the index, W-O, correlates negatively at a substantial level with ability scores. The correlations are, as expected, higher when the index is based on the same test performance as is the measure of ability. This is so for the obvious reason that the ability scores (R and F) are necessarily negatively correlated with W and O--constrained as they are by the number of items in the test, which is constant for all examinees, and also constrained by the fact that R (or F) would necessarily correlate more strongly with W than with O. But even when the correlation is carried out between tests (as in $r_{R_1I_2}$ and $r_{R_2I_1}$, for example), there is a stronger relationship than would be ideal in a study involving guessing and cognitive scores. The same pattern of relationships as that observed between Rights scores and the index is seen in the correlations between Formula scores and the index. The latter correlations, however, are consistently smaller than the former.

It is also noted that the correlations for Group 1 are much smaller than would have been obtained if the students had followed the Rights directions strictly. If they had, there would have been no items omitted or not reached, and the correlation between Rights scores and the index, W-O, would have been -1.00.

Because the various groups tested in the experiment with SAT-verbal were very nearly equal in ability, W-O may be used as an indicator of the extent to which directions affected the tendency to guess. The following table shows the mean Total scores on W-O for groups tested under Rights directions and Formula directions.

Test	Form	Group	Directions	No. of Cases	Mean W-O (Parts 1+2)
SAT-verbal	A	1	Rights	1026	39.28
SAT-verbal	A	4	Formula	1038	31.61
SAT-verbal	B	5	Rights	1040	39.56
SAT-verbal	B	6	Formula	1034	31.35
Chemistry	-	7	Rights	1151	47.24
Chemistry	-	8	Formula	1155	32.04

The tabulations of the "guessing score," W-O, by ability level and ethnic group (Tables 7-12) make it clear that there is more guessing--

at least, more items Wrong, even with the Omits subtracted out--in the lower-scoring groups than in the higher-scoring groups. Again, it should be pointed out that the "guessing score" is necessarily correlated negatively with ability, even on an independent set of items, since the Wrongs component of W-O is very nearly a complement of the Rights, which is expected to show a high correlation across tests. At the same time it should also be pointed out that the Wrongs score, moderated by the Omits, is a reasonable measure of guessing, since, obviously, an item marked incorrectly indicates that the examinee does not know the correct answer to the item and is in fact making either a random or an uneducated guess.

Reading across ethnic groups, it is clear that the Whites do less "guessing" than either the Black or the Black-plus-Hispanic groups, even within strata. This is an observation that, admittedly, may be as much related to the fact that the Wrongs score, which is the heavier component of the W-O guessing index, is necessarily correlated negatively with the Rights score and is evident in these data at least in part because of the breadth of the strata.

When comparisons are made between the two modes of instruction, it is observed that the examinees tested under Rights directions do in fact guess more than do the examinees tested under Formula directions. This difference is not the same at all levels; it is much sharper and clearer for higher-scoring students than for lower-scoring students, suggesting that the lower-scoring students do not follow instructions as well as higher-scoring students or as well as they should, and do not observe the strategies for guessing as wisely as they should.

It is also of some interest to compare the correlations between Part 1 scores and Part 2 scores on Wrongs, Omits and W-0 scores for Groups 1 and 4, who took the test under the same directions on both parts, with the correlations for Groups 2 and 3, which took the test under different directions on the two parts.

Group	Directions on the Two Parts	No. of Cases	Correlations between Parts 1 and 2		
			Wrongs	Omits	W-0
1	Same	1026	.768	.784	.775
2	Different	1068	.686	.246	.538
3	Different	1054	.695	.318	.568
4	Same	1038	.807	.770	.819

These results make it clear that both Wrongs and Omits yield smaller correlations when the directions for the two parts are different than when they are the same, as would be expected. The differences in these correlations are sharper and clearer for Omits than for Wrongs, suggesting that a simple count of the Omits might be an even better indicator of guessing than W-0 because it would not be affected by wrong answers attributable to partial information and misinformation, which tend to impair the clarity of the index.

It would appear that it is not likely that one can develop a satisfactory index of guessing that would be derived solely from the responses to the test itself. Although it is quite reasonable to believe that there should be no correlational relationship between the tendency to guess and cognitive ability in the abstract sense, there is good reason to believe that guessing does affect cognitive test scores and therefore would correlate with them (L. R Tucker, personal communication), as it has in the data cited above. The question, however, is just how strong that relationship should be. The data resulting from this study might conceivably be useful in studying the role of guessing in Rights- and Formula-scored tests, but until we can be more confident of the validity of the guessing index itself, we cannot be confident of the usefulness of the actual data resulting from a study of the correlation of the index with test data.

Effect of Directions and Scoring on Reliability and Parallelism

The design of the study permitted close examination of the correlation between the two parallel sections of the SAT-verbal under all possible combinations of administration and scoring, and also permitted the examination of test-retest and internal consistency estimates of reliability. Finally, it permitted an evaluation of the question whether a change from Formula-scoring to Rights-scoring might not entail so extensive a change in test-taking behavior as to affect the parallelism of the test forms and, consequently, the general applicability of any equating that would be undertaken to make comparable the scores earned under the two types of directions.

Before considering the results of the reliability studies, it should be useful to review certain points that affect their interpretation. First, in interpreting reliability coefficients obtained under Formula directions, the possibility that these coefficients overestimate the reliability of the scores because of the presence of noncognitive variance should be considered. This noncognitive variance arises, according to the Differential Effect Hypothesis, because examinees differ with respect to their willingness to answer questions about which they have useful partial knowledge. On logical grounds, this effect in inflating reliability would be accentuated, rather than diminished, if Rights rather than Formula scoring were used with Formula directions. Second, in interpreting the reliability coefficients for Rights directions, the possibility exists that an analogous effect may arise when, as in this study, there are a substantial number of unanswered items. Here, again, these unanswered items may introduce noncognitive variance into the scores that raises the reliability coefficients artificially. The presence of unanswered items introduces another complication. Suppose that the examinees had followed the instructions to answer every item, and in doing so had engaged in blind guessing. There is reason to believe that the reliability coefficients obtained under such conditions would be lower than those actually found.

If, as seems likely, the effects of these complications on the reliability coefficients are small, the empirical results of this study should be regarded as providing useful but not conclusive evidence on the relative reliability of Rights and Formula scores.

Table 13 provides information on these questions. The first section of the table gives observed-score correlations between Part 1 (45 items) and Part 2 (40 items) under all possible combinations of scoring methods. In spite of the different lengths of the parts of the test, these correlations all represent parallel-forms reliability coefficients and are all comparable insofar as length is concerned, since they all represent the correlation between the 45-item Part 1 and the 40-item Part 2. The italicized figures are the correlations between the results of scoring procedures that were appropriate to the directions actually given in the administrations.

Table 13

Observed-Score Correlations, Reliability Coefficients, and True-Score Correlations between Part-Scores

Form A

Group	No. of Cases	Directions		Observed-Score Correlations[a]				Reliability Coefficients[a]				True-Score Correlations[a]			
		Part 1	Part 2	$r_{R_1R_2}$	$r_{R_1F_2}$	$r_{F_1R_2}$	$r_{F_1F_2}$	$r_{R_1R_1'}$[b]	$r_{F_1F_1'}$[c]	$r_{R_2R_2'}$[b]	$r_{F_2F_2'}$[c]	$r_{\tilde{R}_1\tilde{R}_2}$	$r_{\tilde{R}_1\tilde{F}_2}$	$r_{\tilde{F}_1\tilde{R}_2}$	$r_{\tilde{F}_1\tilde{F}_2}$
1	1026	Rights	Rights	*.822*	.813	.816	.818	*.848*	.838	*.828*	.821	*.981*	.975	.980	.986
2	1068	Rights	Formula	.788	*.798*	.785	.804	*.838*	.830	.809	*.809*	.956	*.970*	.958	.982
3	1054	Formula	Rights	.806	.801	*.810*	.815	.845	*.840*	*.822*	.812	.967	.967	*.975*	.987
4	1038	Formula	Formula	.819	.809	.811	*.824*	.831	*.825*	.803	*.802*	1.003	.991	.997	*1.013*

Form B

Group	No. of Cases	Part 1	Part 2	$r_{R_1R_2}$	$r_{R_1F_2}$	$r_{F_1R_2}$	$r_{F_1F_2}$	$r_{R_1R_1'}$[b]	$r_{F_1F_1'}$[c]	$r_{R_2R_2'}$[b]	$r_{F_2F_2'}$[c]	$r_{\tilde{R}_1\tilde{R}_2}$	$r_{\tilde{R}_1\tilde{F}_2}$	$r_{\tilde{F}_1\tilde{R}_2}$	$r_{\tilde{F}_1\tilde{F}_2}$
5	1040	Rights	Rights	*.820*	.809	.823	.820	*.849*	.837	*.844*	.841	*.968*	.957	.979	.977
6	1034	Formula	Formula	.826	.815	.829	*.837*	.846	*.842*	.842	*.844*	.978	.965	.984	*.992*

[a] Correlations between variables for which the scoring is consistent with the directions appear in italics.

[b] Kuder-Richardson Formula (20) reliability.

[c] Dressel (1940) adaptation of KR (20) reliability.

In general, the correlations are very similar, ranging in Form A from .785 to .824. The two highest correlations in this section of the table are those between the two parts of the test when administered and scored in the same way ($r_{R_1R_2} = .822$ and $r_{F_1F_2} = .824$). The difference between these two correlations is very small and therefore perhaps of little consequence, but it does suggest the possibility that Formula directions and scoring may be slightly more reliable than Rights directions and scoring. This conclusion is supported by the observation that the correlations between Formula scores on the two parts ($r_{F_1F_2}$) are higher, on the average, than any of the other three types of correlations ($r_{R_1R_2}$, $r_{R_1F_2}$, or $r_{F_1R_2}$) even though the method of scoring in some of these instances is inconsistent with the directions. Finally, it is observed that the correlations, $r_{R_1R_2}$ (.82 and $r_{F_1F_2}$ (.824), which are the correlations between two tests administered and scored in the same way, are higher than the other correlations along the main diagonal, $r_{R_1F_2}$ (.798) and $r_{F_1R_2}$ (.810), which are the correlations between tests that were administered and scored in different ways.

The correlations between the parts of Form B also represent a narrow range, from .809 to .837. As in the corresponding first section of the table for Form A, the highest correlation between the parts of the test is that when administered and scored by Formula ($r_{F_1F_2} = .837$), higher than the value, $r_{R_1R_2} = .820$, obtained when Rights directions and scoring were used.

It may also be observed in the data for Form A that the correlations between Parts 1 and 2 are higher when the directions for administration of the two parts are the same (Groups 1 and 4) than when the directions are different (Groups 2 and 3); also that the correlations are higher when the scoring methods are the same (columns 1 and 4) than when the scoring methods are different (columns 2 and 3). However, the differences are not great, indicating that differences in rank ordering of examinees on two tests administered and/or scored in different ways are not much different from the differences in rank ordering when the two tests are administered and scored in the same way. The data for Form B confirm the finding that, on the average, the correlations are slightly higher when the scoring methods for the two parts are the same (columns 1 and 4) than when the scoring methods are different (columns 2 and 3).

The second section of the tables for Form A gives KR (20) reliabilities for each of the two parts of the test, for each of the scoring methods, and for each of the four groups of examinees (and modes of administration). As expected, the reliabilities for Part 1 ($r_{R_1R_1'}$ and $r_{F_1F_1'}$) are higher than the reliabilities for Part 2 ($r_{R_2R_2'}$ and $r_{F_2F_2'}$), since they are based on a slightly longer test section. It is also seen that the KR (20) reliabilities for Rights scores on Part 1 (column 1) are higher than the corresponding reliabilities for Formula scores* (column 2), whether the method of scoring follows the directions for administration or not. This is true for Part 2 also, but not as clearly (see columns 3 and 4).

---

*The reliabilities for Formula scores were calculated using Dressel's (1940) adaptation of KR (20).

The results for Form B are generally consistent with the results for Form A. The KR (20) reliability for Rights is slightly higher than for Formula for Part 1 (.849 vs .842); for Part 2, the reliabilities are equal (.844).

It should be observed that the KR (20) results may overestimate to some extent the reliabilities of the Formula-directed tests because of the slightly greater speededness characteristic of such administrations. (For example, the index of speededness, σ^2_{NR}/σ^2_R, is, almost without exception in these data, higher for Formula-directed administrations than for Rights-directed administrations.) In effect, the reliability of Rights-directed tests, as observed in these data, may actually be undervalued, and the difference in KR (20) reliabilities in favor of Rights-directed tests may well be greater than it appears here.

In general, then, it appears that the two methods of conceptualizing reliability--the correlations between parts 1 and 2 (parallel-forms) and the internal-consistency coefficients (KR (20))--yield contradictory results on this point. However, the differences are quite small, and taken together, the overall results show clearly that the two modes of administration and scoring yield very nearly equal reliabilities. On this basis, it is reasonable to conclude that considerations of test reliability should not be decisive in choosing one method of administration and scoring over the other.

The third section of the table presents evidence on the parallelism of the two conditions of testing and scoring, best indicated by the true-score correlations, .970 and .975, in the second and third columns of the table for Form A. Although these correlations, along with the others in the second and third columns, are generally lower than those in the first and fourth columns--in particular, the values of .981 and 1.013 (essentially, 1.00)--they are sufficiently close to 1.00 to dispel any concerns that they may be measuring substantially different abilities.

Effect of Directions on Equating: Method of Analysis

The principal objective of this study was to evaluate the precision and statistical bias, if any, of equating results obtained by assuming that students perform equally well under Rights directions and Formula directons when Formula scoring is used. Essentially, the analysis methods call for comparing the results obtained by equating Rights scores to Formula scores using standard and "ideal" equating methods with results obtained by equating Rights scores to Formula scores by assuming the Invariance Hypothesis.

In preparation for a description of these methods it will be useful to describe again the method of administering the tests. As set forth earlier in this report, Form A of SAT-verbal was administered to four spiralled groups, each of which was given a different pattern of directions. Form B of SAT-verbal was administered to two spiralled groups (which were also spiralled with the four groups taking Form A), each of which was also given a different pattern of directions. Data for Form B were used principally for confirming the results of some of the analyses for Form A. The Chemistry Test was also administered to two spiralled groups, each of which was also given a different pattern of directions. The sample taking the Chemistry Test was drawn from an entirely different population from the one from which the SAT-verbal (Forms A and B) sample was drawn. Both forms of SAT-verbal were given in two separately-timed 30-minute parts. The Chemistry Test was given in a single 60-minute session.

The directions used for each of the eight groups were as follows:

Group	Directions for Administration	
	Form A	
	Part 1	Part 2
1	Rights	Rights
2	Rights	Formula
3	Formula	Rights
4	Formula	Formula
	Form B	
	Part 1	Part 2
5	Rights	Rights
6	Formula	Formula
	Chemistry	
7	Rights	
8	Formula	

The experimental design permitted the use of five equating methods (but with different subsets of the data), as follows:

1. Spiralling Method. As described earlier in this report, this method calls for distributing the tests in sequence within each room in which the test is administered. As a result of this process, the samples of students taking each form will represent systematic samples of the total group tested.

According to probability theory, each subsample will tend to become increasingly similar to the other subsamples as sample sizes increase. Thus, for large samples it can be assumed that any two subsamples are approximately equal in the abilities measured by the tests to be equated. Scores on two tests are equated by setting equal the means and standard deviations of the samples taking those two tests. The result of the equating is that transformed scores on one test will have the same mean and standard as the observed scores on the other test. (For a fuller discussion of this method, see Angoff (1971, pp. 569-571).)

2. Maximum Likelihood Method. This method calls for administering each of the two tests to be equated to a random sample of students and administering the same link, or anchor, test to all members of both samples. The analytical procedure calls for the estimation of the mean and variance of both tests for the total combined sample, and for setting equal the estimated means and standard deviations for the two tests, as is done in the Spiralling Method. The link test serves to increase the precision of the equating results. This method is described fully by Angoff (1971, pp. 576-579).

The following two methods make use of the Invariance Hypothesis:

3. Invariant Link Method. This method makes use of the same design as that used in the Maximum Likelihood Method. Each group takes one of the two tests that are to be equated. In addition, both groups take the same link test. However, here, one group takes the link test under Rights directions and the other group takes the link test under Formula directions.

Equating is performed by rescoring by Formula the link test taken under Rights directions and assuming that such scores can be treated as interchangeable with Formula scores earned under Formula directions. The analytical method used for treating these data is then identical to that for Maximum Likelihood equating.

4. Invariance Method. In this method, the equating is based on the results of a single test administered to a single group. A test given under Rights directions and scored Rights is also scored by Formula. It is then assumed that the Formula scores so obtained are equivalent to the Formula scores that would have been obtained had Formula directions as well as Formula scoring been employed for that group. The equating procedure then calls for the direct equating of Rights scores to Formula scores for the same individuals by setting equal their means and standard deviations on the two types of scores.

5. Identity Method. Although not a method of equating in the usual sense, it is useful to consider as a criterion, or ideal, the results of an "equating process" which yields a perfectly predictable result. This is one in which a test is "equated" to itself, one which necessarily yields a slope parameter of exactly 1 and an intercept parameter of exactly 0. The advantage of considering this "method of equating" is that the study design permits the equating of scores obtained using a particular type of directions and a particular method of scoring to scores on the same test using the same type of directions and the same method of scoring, but with data based on two independent groups, and the opportunity to compare these

results with the ideal criterion, represented by the Identity Method.

All equating undertaken in the study, except for the auxiliary equating to be described, involved the conversion of:

1) a part score to itself, with the tests for both groups administered and scored by Rights, or with the tests for both groups administered and scored by Formula;

2) a Rights score on a 30-minute part of the test to a Formula score on the same 30-minute part; or

3) a Rights score on the full 60-minute Total test to a Formula score on the full 60-minute Total test

The following enumeration may be helpful in distinguishing among the several types of equatings:

The first of the three types outlined above comprised 12 equatings, as follows:

a) Equating 1R(2) to 1R(1). (Read this: equating Rights scores on Part 1, using data for Sample 2 to Rights scores on Part 1, using data for Sample 1.) The description and the result of this type of equating appear in Table 14, page 78, equations 2 and 3.

b) Equating 2R(3) to 2R(1). Table 14, equations 5 and 6.

c) Equating 1F(3) to 1F(4). Table 14, equations 8 and 9.

d) Equating 2F(2) to 2F(4). Table 14, equations 11 and 12.

These four sets of equatings were each done by the Spiralling and Invariant Link methods and compared with those done by the Identity Method (Table 14, equations 1, 4, 7, and 10).

The second of the three types comprised 18 equatings, and are described as follows:

a) Equating 1R(1+2) to 1F(3+4) by the Spiralling and Invariant Link methods. (Read this: equating Rights scores on Part 1, using data from Sample 1 plus Sample 2 to Formula scores on Part 1, using data from Sample 3 plus Sample 4.) These equatings are found in Table 14, equations 13 and 14.

b) Equating 1R(1) to 1F(3) by the Spiralling and Maximum Likelihood methods. Table 14, equations 16 and 17.

c) Equating 1R(2) to 1F(4) by the Spiralling and Maximum Likelihood methods. Table 14, equations 19 and 20.

d) Equating 1R(1+2) to 1F(1+2) (equation 15), 1R(1) to 1F(1) (equation 18), and 1R(2) to 1F(2) (equation 21), by the Invariance Method. Note that the Invariance Method did not require two samples, as did the Identity, Spiralling, Invariant Link, and Maximum Likelihood methods, but only the one sample, administered under Rights directions.

e) Corresponding to the foregoing 9 equatings were the 9 equatings of 2R to 2F (Table 14, equations 22-30), based on appropriate groups.

The third type of equating outlined on page 73, is further described on page 76.

Additional auxiliary procedures were brought into play for the foregoing two types of equating in order to express all part-score equatings as transformations of: a) raw total Rights scores on Form A to the College Board SAT-verbal scale; b) raw Total Rights scores on Form B of the College

Board SAT-verbal score; and c) raw Total Rights scores on the Chemistry Test to the College Board scale. This process may be described by saying that when a part-score, x, on Form A was to be equated to another part-score, y, on Form A, an attempt was made to express x in terms of Total raw scores on Form A. This was accomplished by applying the mean-and-sigma method to a single sample, rather than to two separate samples, as is ordinarily done in equating one test to another. Specifically, when it was necessary to equate Total scores on Part 1 to Rights scores on Part 1, the data from Sample 1 were used; in our adopted notation TR(1) was equated to 1R(1). Similarly, Total Rights scores were equated to part-scores 2R, 1F, and 2F, as needed, as follows: TR(1) to 2R(1), TR(1) to 1F(4), and TR(1) to 2F(4). In like manner, an attempt was made to express part-score, y, on Form A in terms of Total Formula scores on Form A. This too was accomplished by applying the mean-and-sigma method to a single sample, rather than to two separate samples, and was carried out, as needed, as follows: 1R(4) to TF(4), 2R(4) to TF(4), 1F(4) to TF(4), and 2F(4) to TF(4). Finally, Total Formula scores on Form A were expressed in terms of the College Board scale by the use of conversion parameters for Form A available in file.

The foregoing conversion steps may be summarized in the following diagram:

$$\begin{tikzcd}
\begin{array}{c}\text{Total}\\ \text{Rights}\\ \text{Score}\end{array} \arrow[r, "a"] & \begin{array}{c}\text{Part}\\ \text{Score}\\ (x)\end{array} \arrow[r, "b"] & \begin{array}{c}\text{Part}\\ \text{Score}\\ (y)\end{array} \arrow[r, "c"] & \begin{array}{c}\text{Total}\\ \text{Formula}\\ \text{Score}\end{array} \arrow[r, "d"] & \begin{array}{c}\text{College}\\ \text{Board}\\ \text{Scale}\end{array}
\end{tikzcd}$$

The link of particular interest in this process is the link between one part-score (x) and another (y), or, when analyzed in comparison with the ideal, between a part-score and itself (Link b). The equating link (a) is an auxiliary equating which permits the expression of part-score x in terms of Total Rights score on Form A. The equating link (c) is a second auxiliary equating which permits the expression of part-score y in terms of Total Formula score on Form A. Link d is a set of conversion parameters available in file, developed at the time that Form A was first introduced as an operational SAT, and used to express raw Formula scores on Form A in terms of College Board scaled scores. The application of the succession of links diagrammed above makes it possible to span the links and express all conversions undertaken in the equating of part-scores as conversions of Total Rights scores on Form A to College Board scaled scores. Further, any differences in scaled-score results within each set of three equatings, involving a particular conversion in Link b, are attributable to methodological differences in effecting that link.

The third type of equating outlined on page 73, involving the conversion of Total Rights scores to Total Formula scores, comprised 9 equatings as follows:

a) Equating TR(1) to TF(4) for SAT-verbal, Form A, shown in Table 15, page 85, equations 31 and 32.

b) Equating TR(5) to TF(6) for SAT-verbal, Form B, shown in Table 15, equations 34 and 35.

c) Equating TR(7) to TF(8) for the Chemistry Test, shown in Table 15, equations 37 and 38.

Each of the foregoing equatings was done by the Spiralling and Invariant Link methods.

d) Equating TR(1) to TF(1), TR(5) to TF(5), and TR(7) to TF(7) by the Invariance Method, shown in Table 15, equations 33, 36 and 39. Note again that the Invariance Method did not require two samples, as did the Identity, Spiralling, Invariant Link, and Maximum Likelihood methods, but only the one sample, administered under Rights directions.

In all equating analyses, the examinees' Formula scores were rounded to integers in accordance with the convention used in operational scoring, which is to round upward all scores whose calculated values end in .5. In determining Total scores on Forms A and B of the SAT-verbal, part-scores were rounded to integers before adding them.

Effect of Directions on Equating: Findings

Table 14 describes the essential features of 30 equating sequences relating total Rights scores to Total Formula scores on Form A of SAT-verbal. Results of each equating sequence are expressed on the College Board scale, utilizing file parameters relating Form A Total Formula scores to the standard score scale. Because the equating of Total Rights scores to part-scores (Link a) and the equating of part-scores (Formula) to Total Formula scores (Link c) was performed simply by setting the mean and standard deviation of the designated Total score equal to the mean and standard of the designated part score, the equating method used to establish these links is not specified in Table 14. For the equatings relating part scores,

Table 14

Conversion Parameters Relating Total Rights Score on SAT-Form A to College Board Scale, Determined on Various Data Bases and Various Methods of Equating[a]

	Equating Total Rights Scores to Part Scores		Equating of Part Scores				Equating Part Scores to Total Formula Scores		Parameters		Scaled Score When (X) is:		
Eq. No.	Total Rights Sample	Part and Sample	Part and Sample	Link Test	Part and Sample	Equating Method	Part and Sample	Total Formula Sample	Slope	Intercept	17	40	85
1	1	1R (1)	1R (-)	--	1R (-)	Identity	1R (1)	4	8.1832	70.7859	210	398	766
2	"	"	1R (2)	--	1R (1)	Spiral	"	"	8.4574	60.9408	205	399	780
3	"	"	"	2F	"	Inv.Link	"	"	8.1570	69.5821	208	396	763
4	"	2R (1)	2R (-)	--	2R (-)	Identity	2R (1)	"	8.1838	70.7856	210	398	766
5	"	"	2R (3)	—	2R (1)	Spiral	"	"	8.3065	66.0434	207	398	772
6	"	"	"	1F	"	Inv.Link	"	"	8.2288	70.2317	210	399	770
7	"	1F (4)	1F (-)	—	1F (-)	Identity	1F (4)	"	8.1838	70.7851	210	398	766
8	"	"	1F (3)	--	1F (4)	Spiral	"	"	7.8057	81.4357	214	394	745
9	"	"	"	2F	"	Inv.Link	"	"	8.0561	76.4712	213	399	761
10	"	2F (4)	2F (-)	--	2F (-)	Identity	2F (4)	"	8.1835	70.7836	210	398	766
11	"	"	2F (2)	--	2F (4)	Spiral	"	"	8.0419	75.9928	213	398	760
12	"	"	"	1F	"	Inv.Link	"	"	8.2350	69.9221	210	399	770
13	"	1R (1)	1R (1+2)	—	1F (3+4)	Spiral	1F (4)	"	8.5269	59.9689	205	401	785
14	"	"	"	2F	"	Inv.Link	"	"	8.6454	54.7689	202	401	790
15	"	"	"	—	1F (1+2)	Invariance	"	"	8.7341	51.9801	200	401	794
16	"	"	1R (1)	--	1F (3)	Spiral	"	"	8.5792	59.6247	205	403	789
17	"	"	"	2R	"	Max.Lik.	"	"	8.6650	56.3627	204	403	793
18	"	"	"	—	1F (1)	Invariance	"	"	8.7029	53.2754	201	401	793
19	"	"	1R (2)	—	1F (4)	Spiral	"	"	8.4571	60.9399	205	399	780
20	"	"	"	2F	"	Max.Lik.	"	"	8.5545	57.4419	203	400	785
21	"	"	"	--	1F (2)	Invariance	"	"	8.7654	50.6634	200	401	796
22	"	2R (1)	2R (1+3)	—	2F (2+4)	Spiral	2F (4)	"	8.3190	65.6865	207	398	773
23	"	"	"	1F	"	Inv.Link	"	"	8.5182	60.1132	205	401	784
24	"	"	"	—	2F (1+3)	Invariance	"	"	8.7572	53.5538	202	404	798
25	"	"	2R (1)	--	2F (2)	Spiral	"	"	8.3277	65.4856	207	399	773
26	"	"	"	1R	"	Max.Lik.	"	"	8.5095	59.1411	204	400	782
27	"	"	"	--	2F (1)	Invariance	"	"	8.8050	50.7599	200	403	799
28	"	"	2R (3)	—	2F (4)	Spiral	"	"	8.3067	66.0394	207	398	772
29	"	"	"	1F	"	Max.Lik.	"	"	8.5718	59.2453	205	402	788
30	"	"	"	—	2F (3)	Invariance	"	"	8.7080	56.3567	204	405	797

---

[a] The notations, R and F, in this table refer to the method of scoring used for equating the part, which does not necessarily correspond to the directions under which the part was administered.

however, the table specifies the sample (if any) for each part score and the equating method that was used. In equatings utilizing a link test, both the part designations (1 or 2) and the scoring method (Rights or Formula) are specified.

The first 12 equatings shown in Table 14 all involve equating of a part score to itself. The equating method designated "Identity" shows the results that would be obtained by perfect equating. The equating method designated "Spiralling" shows the results obtained by two essentially random samples that took the designated part. The equating method designated "Invariant Link" uses data on the other separately-timed part of SAT-verbal as an equating section, by assuming that Formula scores for students tested using Rights directions are directly comparable to Formula scores obtained by students tested under Formula directions. The 12 results shown first in Table 14 were obtained by applying each of these methods (Identity, Spiralling, and Invariant Link) to each of 4 part scores (Part 1-Rights, Part 2-Rights, Part 1-Formula, and Part 2-Formula.)

Considering first the results for scaled scores for the 12 equatings, it is clear that there is quite close agreement for a Total Rights score of 40. The range of scaled scores is from 394 to 399. Variation at the chance score level (17 Rights) is from 205 to 214. At the perfect score level (85), the scaled scores range from 745 to 780. Of these 12 equatings, there is one, Equation 8, based on the Spiralling Method, that stands out as clearly different from the rest. If that one equating is set aside, the remaining 11 equatings show a variation of 205 to 213 at the chance score level, 396 to 399 at score 40, and 760 to 780 at score 85.

The variation in maximum scaled scores suggests that it would be useful to compare the slope parameters obtained by the various methods. Figure 1 shows the slopes classified according to the method used to determine them. As would be expected, the 4 equatings based on the fact that the part scores being equated are identical yield slopes that are the same except for rounding error. Results for the Spiralling and for the Invariant Link methods seem to agree reasonably well, on the average, with results of the Identity Method, although the variation of results from one sample to another is noticeably greater for Spiralling than for the Invariant Link Method. This is to be expected on theoretical grounds, in view of the greater sampling error of the Spiralling Method for a given sample size. On the whole, these results may be considered to suggest that the Invariant Link Method should be useful in making the transition from Formula scoring to Rights scoring without disturbing the continuity of the scale.

The remaining 18 equatings shown in Table 14 show relatively little variation in scaled scores corresponding to a Rights score of 40. These scaled scores range from 398 to 405. At the chance score level (17 Rights), the range is from 200 to 207 and at the maximum score level (85), the range is from 772 to 799. Again, it should be useful to consider the possibility of systematic differences in slopes for the various methods. Figure 2 shows the slopes obtained by the various methods. For these 18 equatings, the Invariance Method yields slopes that are consistently high, and Spiralling

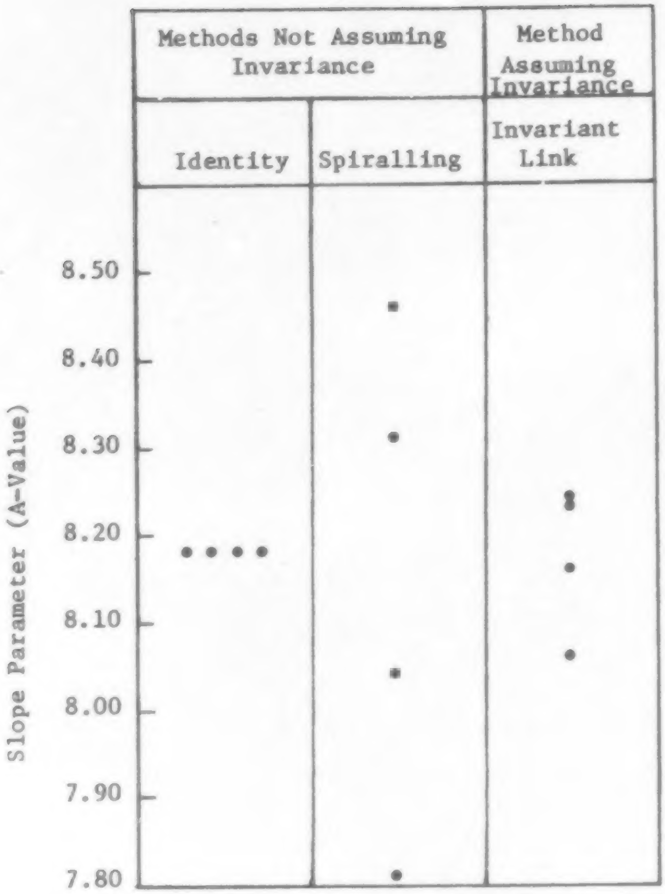

Figure 1. Distribution of Slope Parameters for 12 Equatings Based on Equating Each SAT-verbal Part Score to Itself

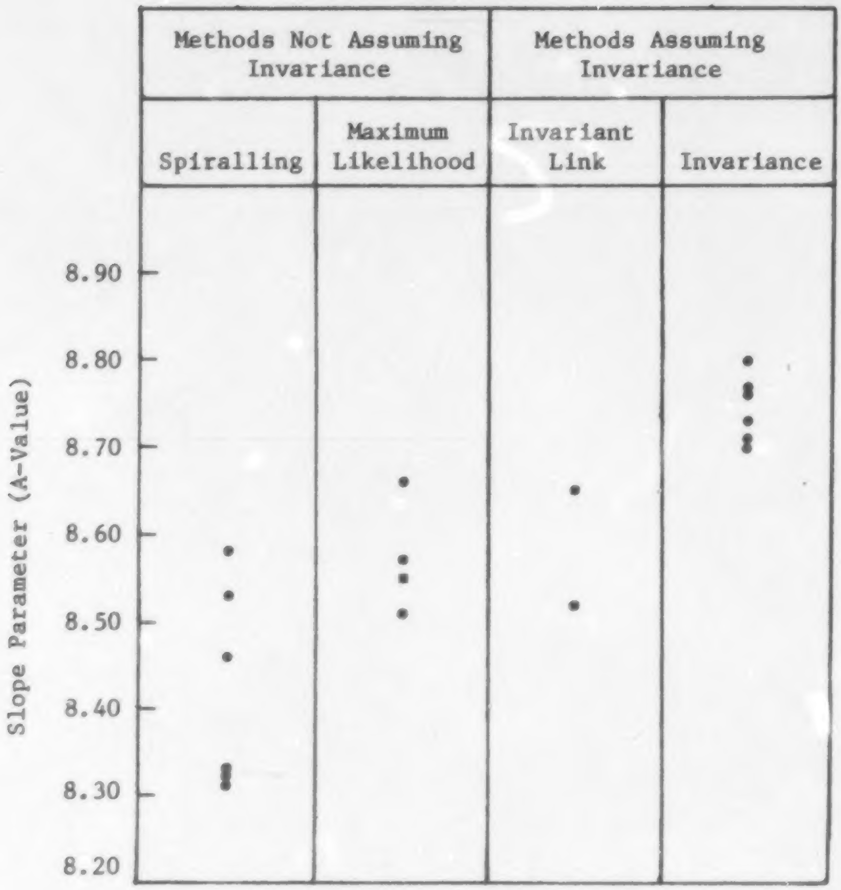

Figure 2. Distribution of Slope Parameters for 18 Equatings Based on Equating Rights and Formula Scores for the Same Part of SAT-verbal

yields slopes that are relatively low. Maximum Likelihood and Invariant Link methods agree reasonably well with each other but yield somewhat higher slopes than those obtained by Spiralling. All six of the slopes for the Invariance Method are relatively consistent, but exceed the highest of the 12 slopes obtained by the other three methods. This result raises some doubt about the practical usefulness of the Invariance Method as a way of relating Rights scores to Formula scores on the same test. The fact that the Invariance Method does not utilize data on the Formula scores earned when the test to be equated is given under Formula-scoring directions may contribute to the difference in results for this method.

On the whole, the results for the slopes for the 30 equations developed for Form A of SAT-verbal suggest that the Invariant Link Method is an acceptable method for relating Rights and Formula scores, but raise questions about the confidence with which the Invariance Method can be used for this purpose. However, it should be noted that the Invariant Link Method may be less satisfactory when applied, under operational conditions, to samples drawn from students attending different test administrations then when applied, as it was in this study, to essentially random equating samples.

The equating analysis of Total scores involved the same three basic methods for all three tests: Forms A and B of SAT-verbal and the Chemistry Test. First, Spiralling was used. Next, the Invariant Link Method was used, with the separately-timed scores on Part 2 as the link test for SAT-verbal and with a subscore based on 25 embedded items that has been used in operational equating as the link test for Chemistry. Finally, Formula scores were obtained for each group that had been tested under Rights directions in order to provide data for the equating by the Invariance Method.

Results for the equating of Total scores are presented in Table 15. Considering first the scaled scores at the mean of the experimental groups, it appears that, although the differences are not large, the methods assuming invariance yield slightly higher scaled scores for raw scores in these comparisons. The differences in scaled scores for maximum raw scores are in the same direction and quite large. At the chance score level, the methods assuming invariance yield somewhat lower scaled scores than does the Spiralling Method. As it turned out, differences in results for the three methods are relatively small for Form B of SAT-verbal.

When slope parameters are considered, the two methods assuming invariance yield similar values for SAT-verbal, and these values are larger than those for the Spiralling Method. For Chemistry, the Invariance Method yields a noticeably steeper slope than does the Invariant Link Method, and the Invariant Link Method yields a noticeably steeper slope than does the Spiralling Method. The fact that the Invariant Link Method yielded a slightly steeper slope than the Invariance Method for Form B of SAT-verbal is inconsistent with the pattern observed for the equatings of part scores. The equating studies show a tendency for equating methods that assume the Invariance Hypothesis to overestimate slightly the slope parameters both

Table 15

Conversion Parameters Relating Total Rights Scores on Forms A and B of SAT-verbal and Total Rights Scores on Chemistry to the College Board Scale, as Determined by Various Methods of Equating

Eq. No.	Test	Form	Group Tested Using Rights Directions	Group Tested Using Formula Directions	Link Test	Equating Method	Parameters		Scaled Score When Raw (Rights) Score is:		
							Slope	Intercept	Chance[a]	Mean[b]	Perfect[c]
31	SAT-verbal	A	1	4	—	Spiral	8.1832	70.7863	210	398	766
32	"	"	"	"	2F	Inv.Link	8.7381	53.1600	202	403	796
33	"	"	"	-	—	Invariance	8.7418	52.6899	201	402	796
34	SAT-verbal	B	5	6	--	Spiral	8.2558	62.7566	203	393	764
35	"	"	"	"	2F	Inv.Link	8.3961	58.7344	201	395	772
36	"	"	"	-	--	Invariance	8.3415	61.4369	203	395	770
37	Chemistry	-	7	8	--	Spiral	7.2568	213.3274	344	460	866[d]
38	"	-	"	"	25 item subscore	Inv.Link	7.6023	205.2797	342	464	889[d]
39	"	-	"	-	—	Invariance	7.9311	193.6605	336	463	907[d]

[a] Chance score is 17 for SAT-verbal; 18 for Chemistry

[b] Mean score is 40 (rounded) for SAT-verbal; 34 (rounded) for Chemistry

[c] Perfect score is 85 for SAT-verbal; 90 for Chemistry

[d] Maximum reported score is 800.

for SAT-verbal and Chemistry, as compared with standard equating methods. There is a tendency for methods assuming the Invariance Hypothesis to yield slightly lower scaled scores at the lower end of the score scale, slightly higher scaled scores for the average student in the study sample, and sufficiently higher scaled scores at the high end of the scale to cause some concern, as compared with standard equating methods. On the other hand, these differences are not alarmingly great. Although they do not warrant a confident assertion that use of the Invariance Hypothesis will permit equating without any danger of a scale discontinuity, they do indicate that the discontinuity, if any, is likely to be within an acceptable tolerance.

Summary and Conclusions

A brief review of the scope of the study and the characteristics of the student samples is relevant to the generalizability of the findings. With respect to scope, the tests administered and the equating methods used need to be considered. Two forms of SAT-verbal and one form of the College Board Chemistry were administered. Because verbal abilities play a prominent role in the large-scale ETS admissions tests, the results for SAT-verbal should be applicable to other verbal tests. The Chemistry Test results provide information on a subject-matter achievement test. The standard equating methods used in the study included the Spiralling Method, the Maximum Likelihood Method, and a method based on equating a test to itself. These methods provided a standard against which equating methods that depended on the use of the Invariance Hypothesis could be compared. With respect to student samples, the study was based on data for high school students who participated in the testing on a voluntary basis. It seems safe to assume that most of these students were reasonably sophisticated about strategies for taking Formula-scored and Rights-scored tests. The sample size for SAT-verbal was over 6,000 and there were more than 1,000 students in each of the six spiralled samples. The sample size for Chemistry was over 2,300, and there were more than 1,150 students in each of the two spiralled groups. These sample sizes should be large enough to provide an adequate basis for investigating the various questions to which the study was addressed.

The data of the study permitted comparisons of Formula and Rights directions with respect to Number of Items Omitted, to Number of Items Not Reached, and to an index of guessing determined by subtracting the Number of Items Omitted from the Number of Wrong Answers. These comparisons yielded differences between the two sets of directions that were in the expected direction but fairly small in magnitude. When students in the samples tested with Form A of SAT-verbal were stratified by ability, results indicated that under the conditions of the study, high-ability students appeared to respond more appropriately to differences in directions than did lower-ability students. Under Rights directions, lower-scoring students tended to omit more items than high-scoring students. Under Formula directions, low-scoring students tended to omit fewer items than higher-scoring students. These findings suggest that special efforts should be made to insure that students answer all items when tests are administered operationally using Rights directions.

The design of the experiment for Form A of SAT-verbal also provided empirical data on two important, long-standing questions for which rigorous data were not hitherto available. First, it was found that when the same tests (Part 1 and Part 2 of SAT-verbal) were administered to one sample under Rights directions and to another, closely comparable sample under Formula directions, the (parallel-forms) correlation between the two parts was higher under Formula directions. However, the difference was very slight. When internal consistency--KR (20)--reliability coefficients were calculated for the tests administered and scored under the two modes, the Rights-administered-and-scored tests were found to be more reliable than the Formula-administered-and-scored tests--in contradiction to the findings for the parallel-forms coefficients. But, again, the difference was very slight, indicating that, in general, the difference in reliability for the two modes of administration and scoring is nonexistent, or at most, of little consequence. Second, it was found that administering Parts 1 and 2 of SAT-verbal with the same directions yielded slightly higher correlations than were obtained if the two parts were administered with different directions. When coefficients were corrected for reliability of measurement, the corrected coefficients were judged to be sufficiently close to 1.00 to justify the assumption of parallelism for purposes of equating.

Results for the effect of differences in directions on mean scores when Formula scoring is used provide some support both for the Differential Effect Hypothesis and for the Invariance Hypothesis. The observed differences for SAT-verbal show slightly higher mean Formula scores (about 5 scaled score points) when Rights directions are used than when Formula directions are used, as predicted by the Differential Effect Hypothesis. On the other hand, the differences are, in general, too small to be statistically significant; and this is just as would be predicted by the Invariance Hypothesis. Moreover, results for Chemistry means are consistent with the Invariance Hypothesis.

The equating studies show that methods that make use of the Invariance Hypothesis, as compared with standard equating methods, might result in a small overestimate of the slope parameters both for SAT-verbal and Chemistry and a fairly large overestimate of scaled scores in the upper portion of the score scale. Thus, although these results do not provide a definitive basis for recommending that the Invariance Hypothesis will permit equating without danger of a scale discontinuity, they do indicate that mean scores should remain reasonably comparable if methods of equating based on the Invariance Hypothesis are used during the period of transition.

A STUDY OF THE GRADUATE MANAGEMENT ADMISSION TEST, BASED ON PROGRAM DATA

Questions Addressed by the Study

The principal purpose for which this study was undertaken was to investigate the effectiveness of several methods of equating scores that had been earned under conditions of Rights directions and scoring to scores earned under conditions of Formula directions and scoring.

In the course of studying the equating methods it was deemed necessary to investigate other related questions:

1. To what extent do the results provide a firm basis for choosing between the Invariance Hypothesis and the Differential Effect Hypothesis?

The Invariance Hypothesis and the Differential Effect Hypothesis differ essentially in their predictions regarding how well students would perform if, instead of choosing to omit certain items when tested under Formula directions, they chose to answer them. The Invariance Hypothesis implies that their performance on the omitted items would be, on the average, neither better nor worse than would be expected by chance. The Differential Effect Hypothesis, on the other hand, implies that their performance on those items would be better, on the average, than would be expected by chance. If the Invariance Hypothesis is supported by the data, Formula scores would remain the same, on the average, whether or not the students chose to omit items about which they had insufficient basis for answering. If, however, the Differential Effect Hypothesis is supported, students who choose to omit certain items when tested under Formula directions would be at a disadvantage in comparison with other students of equal ability who answered all the items.

Although the same student cannot take the same test under both Rights and Formula directions at the same time, it is possible to administer the same test so that one random half of a large group is tested with Rights directions and the other half is tested with Formula directions. The Invariance Hypothesis would predict that the two groups would have virtually equal mean Formula scores; the Differential Effect Hypothesis would predict that the group tested under Rights directions would have a higher mean Formula score than the group tested under Formula directions.

2. To what extent do Formula directions affect the number of items Omitted, number of items Not Reached, and total number of items Not Attempted?

3. When students are stratified on the basis of ability, is there a discernible difference between high and low ability students in the effect of Formula and Rights directions on the average number of items Omitted, Not Reached, or Not Attempted? Do guessing indices defined as "Wrongs minus Omits" or by a formula devised by Ziller provide useful information about guessing tendencies that is not provided by the various indices of nonresponse?

4. To what extent do Formula and Rights directions yield different reliabilities?

5. Is there reason to believe that the assumption of parallelism between a test administered with Rights directions and the same test administered with Formula directions is not warranted?

6. How much confidence can be placed in the Invariance Hypothesis as a basis for equating Rights scores to Formula scores? To what extent does the use of the Invariance Hypothesis result in systematic differences between conversion lines obtained by assuming invariance and corresponding parameters obtained by traditional equating methods?

Study Design

In this experiment the variable section of an operational form of the Graduate Management Admission Test was used to study the effects of Rights and Formula directions. Utilizing these data, taken from the regular administration of the GMAT, offered several advantages over special administrations. In particular, the very large sample size made it possible to study five different item types under the two conditions of administration and to have large samples for each of the ten combinations. With this arrangement, it was possible to use data for the operational part corresponding to each item type in the equating analyses. In addition, conducting the study as part of the GMAT Program provided realistic conditions of motivation and ensured that the sample was reasonably typical of GMAT examinees. On the other hand, conducting the study as part of the program imposed certain restrictions on the study design. In particular, it was considered essential that the experiment be conducted in such a way that it would have no effect on any examinee's score of record, and with the additional condition that no examinee could infer that the test material was experimental and would not cound toward his or her score.

The examinees participating in this experiment undoubtedly anticipated, correctly, that they would be given a Formula-directed test, and therefore it is entirely likely that some of them did not heed the Rights-score directions as closely as they should have. If this was indeed the case, then it would follow that the experimental results have underestimated somewhat the effect of Rights directions on the number of items attempted as compared with what would happen if Rights scoring were used in operational testing. Nevertheless, it seems reasonable that the effects observed in the experiment provide an adequate basis for assessing the impact of different directions on test scores.

The experimental tests were all given as the last of the eight separately-timed sections at the October 1980 administration of GMAT. Each of five item types was administered as a 30-minute separately-timed part, and each of the resulting five experimental tests was administered with Rights directions to one group of students and with Formula directions to a different group of students. (Thus, there were ten different experimental tests.) In order to make the groups taking the different forms comparable in ability level, the tests were spiralled. The order in which the ten tests were packaged and distributed was as follows:

Group	Test	Directions
1	Reading Comprehension	Formula
2	Reading Comprehension	Rights
3	Problem Solving	Formula
4	Problem Solving	Rights
5	Practical Business Judgment	Formula
6	Practical Business Judgment	Rights
7	Data Sufficiency	Formula
8	Data Sufficiency	Rights
9	Sentence Correction	Formula
10	Sentence Correction	Rights

As in the SAT-verbal phase of this study, the same tests were administered under both Rights and Formula directions, and under conditions that permitted the comparison of equivalent groups of examinees who took the tests under one or the other of the two directions. As was noted in the report of the SAT-verbal study, these are the only instances, to our knowledge, in which a study of Rights and Formula directions was designed to permit such comparisons.

Major concerns in planning the test administration were to provide appropriate Rights- and Formula-scoring directions for the experimental tests, in order to ensure that examinees read the directions for those tests and to ensure that they were taken under normal test-taking conditions.

The Supervisor's Manual for the test asked the supervisor to read aloud the following statement just preceding the administration of the experimental tests:

> During the next 30 minutes you are to work only on Section 8. Read the directions at the beginning of Section 8 carefully and answer the questions in accordance with these directions. Turn to Section 8, read the directions and begin to work.

Because the directions were relatively brief, it was decided that examinees could read the directions for their test within the time limits for the test.

The following statement was printed at the beginning of tests given with Formula directions:

> Before answering the questions in this section, please review the following directions, which are the standard directions for GMAT.
>
> Students often ask whether they should guess when they are uncertain about the answer to a question. Your score on this section will be based on the number of questions you answer correctly minus a fraction of the number you answer incorrectly. Therefore, it is improbable that random or haphazard guessing will change your score significantly. If you have some knowledge of a question, you may be able to eliminate one or more of the answer choices as wrong. It is generally to your advantage to answer such questions even though you must guess which of the remaining choices is correct. Remember, however, not to spend too much time on any one question.

The following statement was printed at the beginning of tests given with Rights directions:

> Before answering the questions in this section, please read carefully the following directions, which apply only to this section.
>
> Your score on this test will be based on the number of questions you answer correctly. No deductions will be made for wrong answers. You are advised to use your time effectively and to mark the best answer you can to every question, regardless of how sure you are of the answer you mark.

Although the Formula directions were somewhat longer, it was judged that this difference would be offset by the fact that the directions for these experimental tests were the same directions that had been given to the examinees in previous sections.

Because all examinees who took the experimental tests took the same form of GMAT, and because each of the experimental tests corresponded to one of the parts of the operational test, it was possible to use the operational test part scores in analyzing the experimental test results. The operational test included the following separately-timed parts:

Part*	Content	Number of Items	Number of Minutes
1	Reading Comprehension	25	30
2	Problem Solving	30	40
3	Practical Business Judgment	20	20
4	Data Sufficiency	30	30
5	Practical Business Judgment	20	20
7	Usage	25	15

For purposes of analysis, scores on the two Practical Business Judgment tests were usually combined, and the Usage test was considered to be sufficiently similar to the Sentence Correction test to permit the use of either as a link test for equating scores on the other test.

---

*Part 6 was given as a 30-minute separately-timed part. It was composed of pretest items and did not count in determining the examinee's score, nor was it used in the present study.

Sample Characteristics

The study sample was defined as all examinees for whom scores were reported on the October test and who attempted at least one item on the experimental test. The total sample size was 55,780. The ten subsamples obtained by spiralling ranged in size from 5,408 for Group 5 and 5,409 for Group 10 to 5,739 for Group 1 and 5,738 for Group 6. (The method of packaging and distributing the test books resulted in progressively smaller sample sizes from Group 1 to Group 5 and from Group 6 to Group 10.) Means of GMAT total scores for the ten groups were quite similar, ranging from 473 (Group 6) to 477 (Groups 7 and 8). The overall mean for the total sample was 475, with a standard deviation of 106. By comparison, GMAT candidates tested from November 1975 through July 1978 had a mean score of 461 with a standard deviation of 107.

Results

Effect of Directions on Mean Formula Scores

On logical grounds, the possible differential effect of directions on Formula scores depends on whether examinees would do better than would be expected by chance on items that they answer under Rights directions but do not answer under Formula directions. Although it is anticipated that the means of Rights scores would be clearly higher for those tested under Rights directions than for those tested under Formula directions, it would be expected, under the Invariance Hypothesis, that this difference is caused by random responses to items that normally would not be attempted under Formula directions. Under the Invariance Hypothesis, the difference would be greatly reduced, if not caused to vanish entirely, if a correction for guessing, as is normally applied in Formula scoring, is used. Table 16 provides an opportunity to examine the validity of this hypothesis.

As expected, Table 16 shows that, except for the Practical Business Judgment part, the means of the Rights scores are higher for the groups that received Rights directions than for the groups that received Formula directions. This is not true, however, for the means of the Formula scores. Formula-score means are remarkably similar for the groups receiving the two types of directions on the tests of Reading Comprehension, Problem Solving, and Usage, and reasonably similar on the other two tests. As it happened, the mean Formula score is higher for those tested with Formula directions

Table 16

Descriptive Statistics on Rights and Exact (Unrounded) Formula Scores for GMAT Experimental Tests

Experimental Test	Number of Items	Directions	N	Rights Score[a]		Formula Score[a]	
				Mean	S.D.	Mean	S.D.
Reading Comprehension	29	R	5658	*15.03*	*6.21*	12.35	7.11
Reading Comprehension	29	F	5739	14.68	6.46	*12.38*	*7.17*
Problem Solving	25	R	5501	*9.67*	*4.04*	7.56	4.52
Problem Solving	25	F	5594	9.00	4.02	*7.57*	*4.43*
Practical Business Judgment	32	R	5738	*18.82*	*4.84*	15.67	5.87
Practical Business Judgment	32	F	5408	18.95	4.87	*15.85*	*5.89*
Data Sufficiency	40	R	5590	*16.38*	*5.33*	12.31	5.91
Data Sufficiency	40	F	5657	16.18	5.46	*12.42*	*5.92*
Sentence Correction	30	R	5409	*17.05*	*5.24*	14.44	6.07
Sentence Correction	30	F	5486	16.79	5.40	*14.43*	*6.03*

[a] Means and standard deviations for which the scoring is consistent with the directions appear in italics.

than for those tested with Rights directions in four of the five comparisons. This result is contrary to the hypothesis that examinees would earn better than chance scores on items that they would not answer with Formula directions. On the whole, the results support the hypothesis that Formula scores are invariant with respect to directions.

Effect of Directions on Nonresponse

As described earlier in this report, the entire group of candidates tested in October 1980 was subdivided by spiralling into ten subgroups. Five of these subgroups each took a section of experimental items corresponding in type to one of the five operational sections--Reading Comprehension, Problem Solving, Practical Business Judgment, Data Sufficiency, and Sentence Correction (Sentence Correction to correspond to the operational Usage section)--under Rights directions. The other five groups also each took one of the aforementioned experimental test sections, but under Formula directions. Score intervals were formed in terms of the raw (Formula) scores on the operational test section, and means and standard deviations were tabulated of the number of items Omitted, the number of items Not Reached, and the number of items Not Attempted on the corresponding experimental section for each stratum of students falling in those operational score intervals, as well as for all the students in all strata combined. Means and standard deviations were also tabulated for each of the ten groups of students for the W-O index for guessing and for the Ziller index of guessing, similarly stratified by score on the corresponding operational test section. A summary of the more significant tabulations of these data, as they relate to the effects of directions on the numbers of items attempted appears in Table 17. More detailed tabulations appear in Tables 18-27.

It is recalled that this report distinguishes between unanswered items that precede the last item answered (designated "Omitted") and items that follow the last item answered (designated "Not Reached"). In general, it is reasonable to believe that examinees have considered Omitted items and decided not to answer them. It is also reasonable to believe that examinees have not answered items that are Not Reached because they had too little time to consider them.

Table 17 shows the extent to which groups receiving Rights and Formula directions differed with respect to items Omitted and items Not Reached. Examinees had higher means both on items Omitted and items Not Reached under Formula directions than under Rights directions, as would be expected. The difference in items Omitted is relatively large for Problem Solving. On the other hand, the difference for Practical Business Judgment is trivial. Considering Table 17 as a whole, it appears that, on the average, the effect of directions on the number of items Not Attempted (items Omitted plus items Not Reached) is relatively small.

Tables 18-27 present the same data as in the summary table, Table 17, but in much more detailed form, separately by interval of score on the corresponding section of the operational section. In addition, as already indicated, means and standard deviations of the W-O and Ziller indices are given, similarly by score interval on the corresponding operational section.

Several observations may be made, in Tables 18-27, of the findings in

Table 17

Descriptive Statistics on Number of Items Omitted, Not Reached, and Not Attempted for GMAT Experimental Tests

Experimental Test	Number of Items	Directions	N	Omitted		Not Reached		Not Attempted	
				Mean	S.D.	Mean	S.D.	Mean	S.D.
Reading Comprehension	29	R	5658	1.32	2.67	1.94	3.94	3.26	4.89
Reading Comprehension	29	F	5739	2.44	3.30	2.66	4.35	5.09	5.35
Problem Solving	25	R	5501	4.76	4.84	2.15	3.51	6.90	5.71
Problem Solving	25	F	5594	7.41	4.36	2.90	3.84	10.31	4.41
Practical Business Judgment	32	R	5738	0.44	1.65	0.13	1.10	0.57	2.14
Practical Business Judgment	32	F	5408	0.50	1.72	0.14	1.10	0.64	2.22
Data Sufficiency	40	R	5590	4.31	5.04	3.02	4.78	7.33	6.66
Data Sufficiency	40	F	5657	5.62	5.43	3.18	4.89	8.80	6.69
Sentence Correction	30	R	5409	1.03	2.43	1.47	3.06	2.50	3.92
Sentence Correction	30	F	5486	1.73	2.98	2.05	3.54	3.78	4.49

Table 18

Number of Items Omitted, Not Reached, Not Attempted, and Guessed on the Experimental Section of Reading Comprehension, Given with Rights Directions, Stratified by Formula Scores on the Operational Section of Reading Comprehension

Operational Test Score Interval	No. of Cases	Omitted		Not Reached		Not Attempted		Guessing Index (W-O)		Guessing Index (Ziller)	
		Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
22-25	189	0.29	0.82	0.25	0.89	0.54	1.42	5.14	3.05	0.96	0.13
20-21	529	0.46	1.33	0.56	1.87	1.02	2.43	6.12	3.54	0.96	0.12
18-19	835	0.59	1.54	0.71	2.05	1.30	2.71	7.69	4.12	0.95	0.12
16-17	580	0.93	1.93	0.98	2.36	1.91	3.19	8.34	4.67	0.93	0.14
14-15	879	1.09	2.16	1.48	3.18	2.57	3.96	9.49	5.15	0.93	0.14
12-13	640	1.61	2.73	2.10	3.90	3.71	4.72	9.75	6.03	0.90	0.16
10-11	650	1.78	3.00	2.40	4.17	4.19	5.01	10.92	6.33	0.90	0.16
8- 9	514	1.75	2.87	3.15	4.91	4.90	5.50	11.84	6.41	0.91	0.15
6- 7	339	2.71	4.13	3.38	4.84	6.09	6.09	10.67	8.17	0.86	0.20
4- 5	257	2.07	3.44	4.22	5.62	6.28	6.42	12.69	7.68	0.89	0.18
2- 3	124	2.21	3.48	6.79	7.02	9.00	7.80	12.02	9.05	0.87	0.20
0- 1	122	2.95	5.08	4.69	6.04	7.64	7.49	12.66	10.19	0.87	0.21
Total	5658	1.32	2.67	1.94	3.94	3.26	4.89	9.39	6.12	0.92	0.15

Table 19

Number of Items Omitted, Not Reached, Not Attempted, and Guessed
on the Experimental Section of Reading Comprehension, Given with Formula Directions,
Stratified by Formula Scores on the Operational Section of Reading Comprehension

Operational Test Score Interval	No. of Cases	Omitted		Not Reached		Not Attempted		Guessing Index (W-O)		Guessing Index (Ziller)	
		Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
22-25	186	0.70	1.44	0.72	2.20	1.42	2.92	4.05	3.02	0.90	0.19
20-21	586	0.95	1.82	0.65	1.90	1.60	2.72	5.18	3.46	0.90	0.17
18-19	879	1.23	2.01	0.99	2.44	2.22	3.25	6.11	4.06	0.89	0.16
16-17	568	1.90	2.50	1.86	3.30	3.77	4.15	6.04	4.72	0.85	0.18
14-15	845	2.10	2.75	2.16	3.62	4.26	4.53	7.20	5.21	0.85	0.18
12-13	638	2.93	3.04	3.08	4.27	6.01	4.61	6.92	5.28	0.82	0.17
10-11	681	3.16	3.59	3.21	4.46	6.37	5.15	7.75	6.34	0.82	0.19
8- 9	519	3.44	3.83	4.21	5.00	7.65	5.55	7.90	7.17	0.80	0.20
6- 7	325	4.28	4.41	5.30	5.52	9.58	5.61	6.70	7.30	0.77	0.21
4- 5	253	4.06	4.29	5.16	5.75	9.22	6.23	8.44	7.85	0.80	0.20
2- 3	142	4.63	4.53	6.20	6.23	10.83	6.16	7.68	7.68	0.78	0.19
0- 1	117	4.03	5.21	6.34	7.26	10.37	7.45	9.15	9.38	0.80	0.23
Total	5739	2.44	3.30	2.66	4.35	5.09	5.35	6.79	5.73	0.84	0.19

Table 20

Number of Items Omitted, Not Reached, Not Attempted, and Guessed on the Experimental Section of Problem Solving, Given with Rights Directions, Stratified by Formula Scores on the Operational Section of Problem Solving

Operational Test Score Interval	No. of Cases	Omitted		Not Reached		Not Attempted		Guessing Index (W-0)		Guessing Index (Ziller)	
		Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
27-30	11	2.18	2.25	.45	1.44	2.64	2.46	0.45	3.29	0.59	0.36
24-26	19	2.26	2.45	.53	1.53	2.79	3.02	1.26	4.33	0.58	0.36
21-23	56	2.64	2.77	.50	1.25	3.14	3.02	1.66	4.60	0.67	0.32
18-20	166	2.99	3.09	1.13	2.27	4.12	3.76	2.07	5.61	0.67	0.32
15-17	359	3.80	3.85	1.12	2.21	4.92	4.32	2.18	6.82	0.65	0.33
12-14	713	4.14	4.10	1.57	2.78	5.71	4.68	2.82	7.44	0.67	0.31
9-11	1307	4.49	4.58	2.18	3.46	6.66	5.38	3.46	8.41	0.68	0.31
6- 8	1506	5.24	5.17	2.51	3.84	7.75	5.92	3.53	9.37	0.67	0.31
3- 5	960	5.53	5.31	2.57	3.79	8.10	6.33	4.53	9.94	0.68	0.29
0- 2	404	5.12	5.43	2.42	3.99	7.54	6.57	6.75	10.17	0.73	0.27
Total	5501	4.76	4.84	2.15	3.51	6.90	5.71	3.67	8.86	0.68	0.30

Table 21

Number of Itesm Omitted, Not Reached, Not Attempted, and Guessed on the Experimental Section of Problem Solving, Given with Formula Directions, Stratified by Formula Scores on the Operational Section of Problem Solving

Operational Test Score Interval	No. of Cases	Omitted		Not Reached		Not Attempted		Guessing Index (W-O)		Guessing Index (Ziller)	
		Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
27-30	11	2.64	2.01	0.45	0.89	3.09	2.23	-0.73	1.76	0.45	.29
24-26	32	2.97	2.70	0.38	0.70	3.34	2.84	-0.34	4.01	0.52	.28
21-23	70	3.70	2.91	0.53	1.08	4.23	2.91	-0.23	4.10	0.56	.27
18-20	163	4.75	2.96	1.06	1.80	5.80	3.21	-1.20	4.28	0.49	.27
15-17	357	5.89	3.10	1.64	2.60	7.53	3.04	-1.93	4.36	0.45	.23
12-14	748	6.39	3.55	2.48	3.25	8.87	3.33	-1.87	5.20	0.47	.24
9-11	1346	7.09	3.97	3.14	3.85	10.22	3.73	-1.90	6.01	0.48	.24
6- 8	1518	8.06	4.40	3.33	4.13	11.39	4.13	-2.17	6.69	0.48	.23
3- 5	966	8.81	4.88	3.28	4.21	12.09	4.67	-1.89	7.70	0.50	.24
0- 2	383	8.12	5.09	2.98	4.02	11.10	5.43	1.04	8.87	0.58	.24
Total	5594	7.41	4.36	2.90	3.84	10.31	4.41	1.71	6.54	0.49	.24

Table 22

Number of Items Omitted, Not Reached, Not Attempted, and Guessed on the Experimental Section of Practical Business Judgment, Given with Rights Directions, Stratified by Formula Scores on the Operational Section of Practical Business Judgment

Operational Test Score Interval	No. of Cases	Omitted		Not Reached		Not Attempted		Guessing Index (W-O)		Guessing Index (Ziller)	
		Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
36-40	19	0.00	0.00	0.00	0.00	0.00	0.00	4.58	2.80	1.00	0.00
32-35	124	0.05	0.21	0.02	0.13	0.06	0.25	5.72	2.89	0.99	0.02
28-31	434	0.11	0.73	0.08	1.30	0.19	1.53	8.59	3.83	0.99	0.06
24-27	906	0.16	0.64	0.02	0.18	0.18	0.68	10.10	3.80	0.99	0.06
20-23	1234	0.20	0.94	0.04	0.40	0.24	1.10	11.67	3.75	0.99	0.06
16-19	1321	0.27	0.92	0.05	0.52	0.32	1.13	12.99	3.68	0.99	0.05
12-15	869	0.64	1.75	0.13	0.82	0.77	2.04	13.84	4.44	0.97	0.09
8-11	501	0.91	2.25	0.19	0.95	1.10	2.68	14.78	5.33	0.96	0.11
4- 7	242	1.66	3.27	0.50	2.01	2.16	3.98	16.35	6.65	0.93	0.13
0- 3	88	3.77	6.04	2.53	5.88	6.31	8.53	13.47	11.78	0.84	0.25
Total	5738	0.44	1.65	0.13	1.10	0.57	2.14	12.17	4.89	0.98	0.08

Table 23

Number of Itesm Omitted, Not Reached, Not Attempted, and Guessed on the Experimental Section of Practical Business Judgment, Given with Formula Directions, Stratified by Formula Scores on the Operational Section of Practical Business Judgment

Operational Test Score Interval	No. of Cases	Omitted		Not Reached		Not Attempted		Guessing Index (W-O)		Guessing Inde (Ziller)	
		Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
36-40	17	0.12	0.32	0.00	0.00	0.12	0.32	4.65	2.03	0.98	0.04
32-35	115	0.09	0.34	0.00	0.00	0.09	0.34	6.57	3.77	0.99	0.04
28-31	413	0.10	0.60	0.01	0.08	0.11	0.60	8.23	3.48	0.99	0.05
24-27	897	0.20	0.91	0.07	0.82	0.27	1.45	9.83	3.90	0.98	0.07
20-23	1137	0.21	0.81	0.03	0.45	0.24	1.11	11.48	3.66	0.99	0.05
16-19	1263	0.36	1.13	0.07	0.50	0.43	1.34	12.64	3.80	0.98	0.06
12-15	789	0.59	1.77	0.10	0.63	0.69	1.93	13.69	4.34	0.97	0.08
8-11	453	1.13	2.39	0.21	0.94	1.34	2.69	14.87	5.18	0.95	0.11
4- 7	229	1.97	3.47	0.89	3.46	2.86	4.82	15.31	7.06	0.92	0.14
0- 3	95	3.72	5.63	1.73	4.07	5.44	7.52	13.71	10.58	0.85	0.22
Total	5408	0.50	1.72	0.14	1.10	0.64	2.22	11.91	4.84	0.97	0.08

Table 24

Number of Items Omitted, Not Reached, Not Attempted, and Guessed on the Experimental Section of Data Sufficiency, Given with Rights Directions, Stratified by Formula Scores on the Operational Section of Data Sufficiency

Operational Test Score Interval	No. of Cases	Omitted		Not Reached		Not Attempted		Guessing Index (W-0)		Guessing Index (Ziller)	
		Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
27-30	19	1.68	2.64	0.53	1.39	2.21	3.09	8.05	5.30	0.88	0.17
24-26	143	1.91	2.80	1.69	3.25	3.59	4.48	8.74	5.32	0.87	0.17
21-23	261	2.31	3.07	2.41	3.95	4.71	4.91	9.90	5.83	0.87	0.16
18-20	591	2.93	3.77	2.60	4.09	5.53	5.35	10.45	6.83	0.85	0.17
15-17	658	3.68	4.09	3.25	4.93	6.92	6.15	11.00	7.83	0.83	0.18
12-14	889	4.19	4.70	3.24	4.90	7.43	6.44	11.25	8.65	0.82	0.19
9-11	1033	4.79	5.12	3.23	4.96	8.02	6.83	11.82	9.46	0.81	0.19
6- 8	879	5.33	5.75	3.29	5.07	8.62	7.04	12.02	10.44	0.80	0.20
3- 5	662	5.11	5.81	2.82	4.69	7.94	7.21	14.29	11.08	0.82	0.20
0- 2	455	5.09	5.79	2.96	5.09	8.05	7.39	16.03	11.02	0.84	0.18
Total	5590	4.31	5.04	3.02	4.78	7.33	6.66	11.97	9.35	0.83	0.19

Table 25

Number of Items Omitted, Not Reached, Not Attempted, and Guessed on the Experimental Section of Data Sufficiency, Given with Formula Directions, Stratified by Formula Scores on the Operational Section of Data Sufficiency

Operational Test Score Interval	No. of Cases	Omitted		Not Reached		Not Attempted		Guessing Index (W-O)		Guessing Index (Ziller)	
		Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
27-30	24	1.29	1.99	1.83	3.48	3.13	3.55	7.00	3.20	0.91	0.13
24-26	154	1.64	2.61	1.42	2.99	3.06	3.87	9.42	4.49	0.90	0.14
21-23	276	2.96	3.37	2.82	4.27	5.78	5.03	8.57	5.98	0.84	0.18
18-20	575	2.94	3.33	2.80	4.21	5.74	5.24	10.03	5.93	0.85	0.16
15-17	687	4.86	4.71	3.49	4.88	8.35	6.16	8.60	8.08	0.78	0.19
12-14	846	5.48	4.87	3.64	[illegible]	9.13	6.24	8.51	8.32	0.76	0.19
9-11	1064	5.70	5.17	3.44	5.14	9.15	6.57	9.70	9.09	0.77	0.19
6- 8	816	7.02	5.69	3.32	5.05	10.34	6.65	8.95	9.81	0.74	0.19
3- 5	726	7.33	6.17	3.15	5.25	10.48	7.35	9.87	10.88	0.75	0.20
0- 2	489	8.04	6.80	2.44	4.33	10.47	7.53	11.26	12.26	0.75	0.21
Total	5657	5.62	5.43	3.18	4.89	8.80	6.69	9.39	9.08	0.78	0.19

Table 26

Number of Items Omitted, Not Reached, Not Attempted, and Guessed on the Experimental Section of Sentence Correction, Given with Rights Directions, Stratified by Formula Scores on the Operational Section of Usage

Operational Test Score Interval	No. of Cases	Omitted		Not Reached		Not Attempted		Guessing Index (W-O)		Guessing Index (Ziller)	
		Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
22-25	140	0.21	0.72	0.14	0.63	0.36	0.96	4.85	2.78	0.95	0.18
20-21	274	0.27	0.85	0.41	1.46	0.68	1.74	5.93	2.87	0.97	0.09
18-19	448	0.39	1.09	0.62	1.82	1.01	2.23	6.81	3.27	0.96	0.11
16-17	333	0.67	1.84	1.02	2.28	1.68	2.89	7.20	3.95	0.94	0.14
14-15	628	0.51	1.40	1.07	2.43	1.58	2.86	8.76	3.95	0.96	0.10
12-13	589	0.93	1.93	1.31	2.78	2.24	3.38	9.08	4.35	0.94	0.12
10-11	748	0.87	1.93	1.46	3.10	2.33	3.61	9.78	4.53	0.94	0.12
8- 9	778	1.11	2.35	1.65	3.26	2.76	4.02	10.99	5.25	0.93	0.13
6- 7	545	1.67	3.18	2.12	3.54	3.79	4.62	10.40	6.04	0.91	0.16
4- 5	502	1.81	3.49	2.18	3.70	3.99	4.85	11.36	6.60	0.91	0.17
2- 3	266	1.87	3.63	2.72	4.03	4.59	5.11	11.74	6.94	0.91	0.16
0- 1	158	2.22	4.08	2.59	4.19	4.81	5.60	12.73	8.38	0.89	0.19
Total	5409	1.03	2.43	1.47	3.06	2.50	3.92	9.42	5.36	0.93	0.14

Table 27

Number of Items Omitted, Not Reached, Not Attempted, and Guessed on the Experimental Section of Sentence Correction, Given with Formula Directions, Stratified by Formula Scores on the Operational Section of Usage

Operational Test Score Interval	No. of Cases	Omitted		Not Reached		Not Attempted		Guessing Index (W-O)		Guessing Index (Ziller)	
		Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
22-25	121	0.49	1.20	0.44	1.34	0.93	1.88	4.39	2.72	0.94	0.14
20-21	279	0.41	1.07	0.61	1.75	1.02	2.09	5.25	2.92	0.94	0.16
18-19	472	0.75	1.67	0.85	2.25	1.60	2.94	6.10	3.60	0.92	0.16
16-17	353	1.05	1.92	1.27	2.65	2.32	3.16	6.53	3.68	0.91	0.14
14-15	648	0.84	1.73	1.08	2.41	1.92	3.01	7.79	4.00	0.93	0.13
12-13	588	1.68	2.56	1.62	2.94	3.30	3.70	7.45	4.56	0.88	0.16
10-11	822	1.70	2.87	1.98	3.32	3.68	4.21	8.31	5.21	0.89	0.17
8- 9	741	1.85	2.78	2.61	3.76	4.46	4.42	8.77	5.29	0.88	0.16
6- 7	555	2.79	3.90	2.94	3.98	5.74	4.87	8.05	6.36	0.85	0.19
4- 5	483	2.92	3.84	3.43	4.54	6.35	5.14	8.15	6.42	0.84	0.19
2- 3	276	3.12	4.24	4.06	5.03	7.18	5.77	8.76	6.81	0.84	0.19
0- 1	148	3.08	4.50	3.80	4.53	6.89	5.62	10.15	8.10	0.85	0.20
Total	5486	1.73	2.98	2.05	3.54	3.78	4.49	7.70	5.26	0.89	0.17

the columns of numbers of items Omitted. One, as expected, there is a decline in the mean number of items Omitted as a function of score on the corresponding operational test; the higher the score on the operational test, the fewer the Omits on the corresponding experimental section. Two, also as expected (and as pointed out above), there are more Omits for those taking the test under Formula directions than for those taking the test under Rights directions. This difference is clearly in evidence on all of the five experimental tests except for Practical Business Judgment, in which the difference, while still in favor of those tested under Formula directions, is very small indeed. Three, the progressions of mean Omits for the two types of directions track each other very closely in most of these five tests, especially in the case of Practical Business Judgment, where they follow almost precisely the same levels as well as the same patterns. Four, in all five test sections the differences in the mean number of items Omitted for the two types of directions become smaller with increases in ability; the higher the score on the operational test, the smaller is the difference in the mean Omits for candidates tested under Formula and Rights directions. This last finding, it is noted, is in conflict with the finding in the SAT phase of the study, in which it was observed that differences in omitting behavior were more pronounced at higher levels of ability than at lower levels of ability, not less pronounced. Whether the difference between the two studies is a function of the age and level of sophistication of the students is a matter for speculation.

The results shown by the tabulations of the number of items Not Reached match directly those shown by the tabulations of the number of items Omitted. One, the number of items Not Reached declines with score on the corresponding operational section of the GMAT. This too is to be expected, since the count of Not Reached (NR) is often taken as a measure of speededness, and speededness is expected to correlate negatively with ability. Two, the tests administered under Formula directions show higher average NR counts than tests administered under Rights directions. This too is expected; Formula-directed tests are likely to take more time per item and to result in greater numbers of items Not Reached than Rights-directed tests. In addition, because it is to the student's advantage under Rights directions to answer every item, some examinees undoubtedly used blind guessing or superficial considerations in answering items near the end of the tests. Three, also as in the study of Omits, the progression of mean NR counts for Formula-directed tests and the progression of NR counts for Rights-directed tests track each other surprisingly closely, especially so (again) for the test of Practical Business Judgment. Four, the differences between mean NR counts for Formula-directed tests and mean NR counts of Rights-directed tests decline as a function of increasing ability, as measured by the score on the corresponding operational test section.

The tabulations of the mean number of items Not Attempted (NA) show the same pattern as shown by the tabulations of the numbers of items Omitted and Not Reached. This too is expected, inasmuch as the NA count is the simple sum of the counts of Omitted and Not Reached items. As in the case of its component counts, (a) the number of items Not Attempted declines as the score on the corresponding operational test section rises; (b) the NA count is clearly and consistently higher for Formula-directed tests than for Rights-directed tests _except_ in the case of Practical Business Judgment, for which the two trends are very close and almost indistinguishable; (c) the progression of the decline for the NA count in the Rights-administered tests tracks very closely the progression of the decline for the NA count in the Formula-administered tests, and as was just pointed out, they are virtually indistinguishable in the case of the Practical Business Judgment test; and (d) the difference in the NA count decreases with ability, as measured by the operational section score.

Effect of Directions on Guessing

In the present phase of the study, two indices of guessing were studied. One of these indices is the index, W-O (the number of items answered incorrectly minus the number of items omitted), examined in the study of the SAT. The other is the index advanced by Ziller (1957):

$$Z = \frac{[k/(k-1)]\ W}{[k/(k-1)]\ W + NA},$$

where k = no. of response options per item,

W = no. of items answered incorrectly, and

NA = no. of items Not Attempted (Omitted plus Not Reached).

Both of these indices are offered for consideration because both appear to benefit from a defensible rationale. At the same time, it should be pointed out that both suffer from certain deficiencies. Both indices, it is observed, are derived from the responses, and nonresponses, to the items in the test. The justification for both indices is the justification described in the report of the SAT phase of this study: that the number of Wrongs includes all items for which it can safely be assumed that the student had less than complete knowledge. Assuming that he (she) made no clerical errors in responding, it is presumed that anyone who responds without complete knowledge does so with at least some degree of guesswork--except, perhaps, for those students who respond with confidence but with incorrect knowledge or information.

As suggested earlier in this report, it is reasonable to believe that there should be no correlational relationship between the tendency to guess, when taken as a personality trait, and cognitive ability in the abstract sense. On the other hand, there is good reason to believe that the act of guessing does affect the test score and therefore should correlate with it (L. R Tucker, personal communication). The question remains, to what extent should there be such a correlation, and should the correlation be positive or negative? Related to this question might be the question, should we be able to anticipate the nature of the correlation from the nature of the index itself?

As described above, the rationale for the index W-O is that W includes all items for which it may be assumed that the student had less than complete knowledge, and it is presumed that, except for the instances of confident, but incorrect, responses, the student who responds without complete knowledge does so with some degree of guesswork. The subtraction of Omits is introduced as evidence of a tendency not to guess. It might be argued that there is a deficiency in the W-0 index, arising principally from the fact that it does not control for "opportunity." That is, leaving aside the NR count--which is taken to be the number of items that the student does not have time to consider, and therefore does not represent either a decision to guess or a decision not to guess--the index W-0 is largely a function of score level. Thus, high-ability students will necessarily earn a low W-0 index of guessing, and this can be predicted from an examination of the index itself; in effect, the student's level of ability coerces the numerical value of the guessing index. On the other hand, if we are searching for an index of guessing behavior on the test, as we are, then the W-0 index becomes far more attractive. High-ability students have a low W-0 value because they do not guess. For obvious reasons they do not have to guess. The index, then, is admittedly not a measure of their general propensity to guess; it is a measure of the amount of guessing they actually do in the course of taking that test.

There is one adjustment that might have been introduced into the W-0 index, but was not. This is to increase the W component by the factor k/(k-1) (where k = number of response options per item), to account for the fact that some of the student's guessing actually resulted in correct responses. Although the factor k/(k-1) may well constitute an overcorrection--because many correct responses are the result of partial information and therefore only partial guessing--it is also clear that the unadjusted index, W-0, is a slight underestimate of the actual guessing behavior. On the other hand, it is probably not enough of an underestimate to change the results of this study in any significant way.

The Ziller index, it is noted, does contain the factor k/(k-1), and expresses the index of guessing as a proportion, in which the numerator represents the number of times the student did guess, and in which the denominator represents an "ignorance" score, a score representing the maximum number of items at which the student might have guessed. Indeed, the denominator is algebraically identical to the total number of items minus the conventional Formula score, $R - \frac{W}{k-1}$. In this sense, the index is properly characterized as an attempt to describe a general tendency on the part of the student, transcending his (her) actual behavior on the particular test. This being the case, one would expect--and the data cited below confirm this expectation--that the Ziller index will show a lower correlation with score level than the W-0 index.

There is no fundamental conflict between the two indices; they are intended to express different types of measures. Nevertheless, it may also be in order to suggest some possible alterations in the Ziller index, as we did with the W-0 index. First, it would be useful to include in the numerator a subtraction for Omits, as is done in the W-0 index, because the omission of an item indicates a tendency not to guess. (Note that this change would make the numerator in the Ziller index identical to the W-0 index, once the W in that index is weighted by the factor, k/(k-1).) Second, in consideration of the intent of the denominator, which is to express the number of opportunities to guess, it would be preferable to confine the nonresponses in the second term to Omits only, on the basis that the NR items are those that the student has never had time to consider. Thus, the suggested revision of the Ziller index might be the following:

$$Z' = \frac{[k/(k-1)]\ W - O}{[k/(k-1)]\ W + O}\ .$$

It was expected in this phase of the study, as in the SAT phase of the study, that the W-O index would correlate negatively with score level because W, the dominant factor in the index, would certainly correlate negatively with score level. It was not known, however, how the Ziller index would correlate, and whether the correlation might yield any insights regarding the characteristics of Rights and Formula directions. Table 28 below is designed to offer some information in these respects. The first section of this table gives correlations of Wrongs, Omits, W-O, and Ziller, earned on an experimental section (when administered and scored by Rights and when administered and scored by Formula) with the corresponding Formula-scored operational section of the test. The second section of the table gives the correlations of the same four variables with the experimental test score from which the variables themselves were derived. The correlations in this section of the table describe relationships between variables that are based, in part, on the same data. As expected, both the Wrongs scores and the Omits scores are without exception negatively correlated with both the operational and the experimental test sections. Also as expected, the W-O index is, with only two exceptions, negatively correlated with the operational scores and with the scores on the experimental

Table 28

Correlations of Wrongs, Omits, and Two Indices of Guessing on Experimental Tests with Scores on Operational and Experimental Tests

			Correlations of Responses on Experimental Tests with:							
			Operational Test Score[a]				Experimental Test Score[a]			
Experimental Test	Directions	No. of Cases	Wrongs	Omits	W-O	Ziller	Wrongs	Omits	W-O	Ziller
Reading Comp.	Rights	5658	-.497	-.249	-.310	.175	-.644	-.352	-.388	.258
	Formula	5739	-.451	-.332	-.164	.204	-.675	-.391	-.305	.212
Prob. Solving	Rights	5501	-.330	-.133	-.117	-.048	-.234	-.374	.069	.159
	Formula	5594	-.347	-.257	-.026	-.064	-.531	-.290	-.109	-.162
Pract. Business Judgment	Rights	5738	-.562	-.259	-.432	.213	-.898	-.327	-.718	.243
	Formula	5408	-.568	-.284	-.422	.232	-.926	-.294	-.749	.220
Data Sufficiency	Rights	5590	-.433	-.179	-.172	.067	-.287	-.342	.006	.245
	Formula	5657	-.393	-.301	-.051	.175	-.443	-.311	-.075	.175
Sentence Corr.[b]	Rights	5409	-.525	-.213	-.343	.135	-.687	-.361	-.412	.248
	Formula	5486	-.453	-.272	-.192	.181	-.684	-.369	-.314	.226

[a] Scores on the operational tests are uniformly Formula scores. Scores on the experimental tests are consistent with the directions: Rights scores with Rights directions and Formula scores with Formula directions.

[b] The operational score corresponding to the experimental Sentence Correction section is Usage.

tests. The range of the correlations with the operational scores extends from -.026 to -.432. The range of the correlations with the experimental scores, leaving aside the Practical Business Judgment test for the moment, extends from -.412 to +.069. The correlations of W-0 with the experimental Practical Business Judgment test are extremely high negative in comparison with the others (-.718 and -.749) undoubtedly because the means and standard deviations of Omits shown in Tables 22 and 23 are so small that W-0 becomes virtually equal to W. These correlations are indeed not much lower than the correlations of the Wrongs with the experimental test scores (-.898 and -.926). Referring again to the column of correlations of W-0 with the operational test scores, we see that the correlations with Practical Business Judgment are again high negative. Reference to the operational test analysis for that form confirms that that operational test also had very few Omits, again suggesting that the correlation is to a considerable extent the correlation between complementary scores. Here too, the correlations of the W-0 socres with the operational test scores (-.432 and -.422) are not much lower than the correlations of the Wrongs scores with the operational test scores (-.562 and -.568).

The correlations of the Ziller index with the operational test scores and with the experimental test scores are generally positive and smaller, in absolute size, than the corresponding correlations for the W-0 index. This is expected, as indicated earlier, because, unlike the W-0 index, which is a direct function of the student's behavior on the test, the Ziller index, in expressing the "amount of guessing" as a proportion of the "opportunity for guessing," attempts to achieve a measure of the more general "tendency to guess."

Returning to Tables 18-27, it is noted that the same tabulations as those made for the counts of items Omitted, Not Reached, and Not Attempted were made for the two guessing indices. As in the results of the preceding three counts, the curve of the W-0 index is seen to follow a declining pattern, a pattern which is expected on theoretical grounds, and also expected in view of the observed negative correlation of W-0 and test scores. Second, as in the tabulations of 0, NR, and NA, the level of guessing is clearly greater for Rights-directed tests than for Formula-directed tests, except for the test of Practical Business Judgment, where the amount of guessing, as measured by the W-0 index, is virtually the same for the two modes of directions. Third, the curves of the two sets of W-0 means track each other very closely, especially, again, for the test of Practical Business Judgment. Finally, the mean values of the index approach each other as one moves up the scale of ability, rather than diverge from each other, as was observed in the SAT phase of the study.

The Ziller index of guessing behaves very much as does the W-0 index, except that it tends to rise, rather than decline, with the score on the corresponding operational section. But like the W-0 index, it shows generally higher mean values for Rights- than for Formula-directed areas, except for the Practical Business Judgment test where the means are very similar; and it shows generally the same fluctuations in the progression of its means with categories of score on the operational test. Unlike the

W-0 index, there is no clear tendency for the two sets of means either to converge or diverge as a function of score on the operational test.

Effect of Directions and Scoring on Reliability and Parallelism

Because the GMAT study was performed as part of an operational test administration, it was not possible to vary the directions for the separately-timed parts as was done in the SAT-verbal experiment. However, the data provided by the GMAT study do permit the comparison of parallel-forms correlations between tests both of which were administered and scored with the same (Formula) directions with parallel-forms correlations between tests administered and scored with different directions. They also permit the comparison of KR (20) reliabilities under the two conditions of administration and scoring.* Finally, they permit the comparison of true-score correlations between two parallel tests in evaluating the question whether a change from one type of administration and scoring to another might not cause an extensive shift in the nature of the ability measured.

It is recalled that the administration that permitted the tabulation of these data was a regular administration of the GMAT, in October 1980, at which time certain conditions had to be met in order to provide reportable scores for the students sitting for the tests at that administration. Clearly, the operational scores--the scores of record--had to be earned under Formula directions. Second, the tests administered under different directions were composed of items that were being pretested for possible operational use. Such items cannot be expected to be of the uniformly high quality that characterizes the items in the operational forms of the GMAT. Third, one of the five experimental tests used in the study, Sentence Correction, was not parallel to the corresponding Usage test, as would be ideal in a study of this sort. These foregoing considerations are relevant to the interpretation of the results shown in Tables 28 and 29.

---

*For a discussion of certain logical considerations in interpreting these reliability estimates, see pages 63 and 64.

Observed-score correlations of the five experimental tests with the corresponding test material of the six operational tests are shown in the first section of Table 29. (In this table, correlations for which directions and scoring are consistent with each other are italicized.) For each item type, correlations are shown for both Rights and Formula directions and for both Rights and Formula scoring. In general, the four correlations for a given combination of operational test and experimental test are remarkably similar. On the other hand, there is considerable variation in correlations among the sets of correlations for different tests, indicating that the size of the correlations is more a function of the particular test than of directions or scoring. Within each test, however, with one minor exception (Practical Business Judgment--Section 3), the correlations between the operational tests administered and scored by Formula and the experimental tests administered and scored by Formula ($r_{F_o} r_{F_e}$) are slightly higher than the correlations between operational tests administered and scored by Formula and experimental tests administered and scored by Rights ($r_{F_oR_e}$). Because of the constraints on the administrations, discussed above, it was not possible to observe the comparison between tests

Table 29

Observed-Score Correlations, Reliabilities, and True-Score Correlations between Experimental and Operational Test Sections

Operational Test	Directions for Experimental Tests	N	Observed-Score Correlations of Experimental Tests with Corresponding Operational Test[a,b]		Reliability Coefficients[a,b]				True-Score Correlations of Experimental Tests with Operational Test[a,b]	
			$r_{F_oR_e}$	$r_{F_oF_e}$	$r_{R_oR_o}'^{c}$	$r_{F_oF_o}'^{d}$	$r_{R_eR_e}'^{c}$	$r_{F_eF_e}'^{d}$	$r_{\tilde{F}_o\tilde{R}_e}$	$r_{\tilde{F}_o\tilde{F}_e}$
Reading Comprehension	R	5658	*.728*	.726	.793	*.766*	*.854*	.836	*.900*	.907
	F	5739	.733	*.732*	.796	*.771*	.868	*.848*	.896	*.905*
Problem Solving	R	5501	*.687*	.709	.754	*.742*	*.719*	.686	*.940*	.994
	F	5594	.735	*.740*	.760	*.752*	.736	*.720*	.988	*1.006*
Practical Business Judgment-Section 3	R	5738	*.593*	.589	.660	*.625*	*.748*	.736	*.868*	.869
	F	5408	.595	*.592*	.662	*.631*	.752	*.737*	.864	*.868*
Practical Business Judgment-Section 5	R	5738	*.543*	.536	.619	*.588*	*.748*[e]	.736[e]	*.818*	.816
	F	5408	.559	*.553*	.614	*.586*	.752[e]	*.737*[e]	.842	*.840*
Data Sufficiency	R	5590	*.676*	.716	.822	*.812*	*.751*	.710	*.866*	.944
	F	5657	.695	*.731*	.829	*.819*	.768	*.722*	.877	*.950*
Usage	R	5409	*.670*	.677	.782	*.755*	*.782*	.758	*.873*	.895
	F	5486	.672	*.677*	.781	*.750*	.796	*.763*	.869	*.894*

[a] Operational (Formula) scores used in this table are unrounded. Also, if scores are found to be negative, they are used as negative.

[b] Correlations between variables for which the scoring is consistent with the directions appear in italics.

[c] Kuder-Richardson Formula (20) reliability.

[d] Dressel (1940) adaptation of KR (20) reliability.

[e] Note that these reliabilities are identical to the corresponding reliabilities shown for Practical Business Judgment-Section 3. Although there were two operational sections of this type, Section 3 and Section 5, there was only one such experimental section.

that are both administered and scored by Rights ($r_{R_oR_e}$) and tests that are administered and scored in different ways--e.g., $r_{F_oR_e}$ or $r_{R_oF_e}$.

The second section of Table 29 gives KR (20) reliabilities for Rights and Formula scores on both the operational and experimental sections of the test. In this section of the Table, as in the first section, the italicized numbers apply to reliabilities calculated under scoring conditions consistent with the directions used in administering the tests. Several comments may be made on these data. First, it is seen that the reliabilities of four of the five experimental tests when administered and scored Rights are higher than the reliabilities of the same tests when administered and scored by Formula. The differences, however, are relatively small, the largest being .029. Second, Rights scores yield higher reliability coefficients than do Formula scores in all 12 comparisons for the operational tests and in all 10 comparisons for the experimental tests. In the 22 comparisons, the differences range from .008 to .046, with a median of .026. These findings suggest that Rights scoring provides more reliable scores than does Formula scoring. On the other hand, the results for the experimental tests indicate that Formula directions yield higher reliability coefficients than do Rights directions, for both methods of scoring. Thus, when directions are compared, Formula directions yield higher reliabilities, and when scoring methods are compared, Rights scoring yields higher reliabilities. On the whole, then, the internal consistency reliability results do not provide an unequivocal basis for preferring Rights-directions-and-scoring to Formula-directions-and-scoring.

The data also permit a comparison of the reliability coefficients of the operational and experimental tests. The table below gives the numbers of items and time limits for each operational and experimental test.

Numbers of Items and Time Limits for the Operational and Experimental Tests

Test Section		Operational		Experimental	
		No. of Items	Time (Mins.)	No. of Items	Time (Mins.)
Reading Comprehension		25	30	29	30
Problem Solving		30	40	25	30
Practical Business Judgment	Section 3: Section 5:	20 20	20 20	32	30
Data Sufficiency		30	30	40	30
Usage		25	15	30*	30

Results for reliability coefficients of Formula scores shown in Table 29 indicate that in all eight comparisons for which time limits of the experimental and operational tests were different (Problem Solving, Practical Business Judgment, and Usage/Sentence Correction), the test having the longer time limit was more reliable. In the four comparisons based on operational and experimental tests having equal time limits, the experimental test had the higher reliability for Reading Comprehension and the operational test had the higher reliability for Data Sufficiency.

The third section of Table 29 gives estimates of true-score correlations between each operational test and its corresponding experimental test, administered and scored under each of the two modes, Formula and Rights. The study design provided internal-consistency estimates of reliability for the operational and experimental tests. To the extent that the observed-score correlations involved sources of error not present in the KR (20) and Dressel determinations of reliability, and to the extent that these determinations underestimate the reliability, the estimates of the true-score correlations are too high. With the exception of the Problem Solving test the estimated correlations of true scores are noticeably lower than the value of 1.00 that would be expected for parallel tests. This divergence from parallelism is probably attributable, in large part, to the fact that pretests rather than operational tests were used in the experiment. The fact that the SAT-verbal parts in the first phase of this study yielded coefficients nearer to 1.00 would be consistent with this interpretation because the SAT-verbal tests were operational forms. It is plausible, also, that variation in results for different item types may arise because it is more difficult to construct strictly parallel tests for some item types than for others.

---

*The experimental section corresponding to the operational Usage section consisted of Sentence Correction items.

Despite their limitations, the data provide some useful comparisons of true-score correlations obtained when directions and scoring are the same for the two tests with corresponding correlations obtained when directions and scoring differ from one test to the other. For five of the six tests, the same directions yield higher true-score correlations than do different directions; for the remaining comparisons, the correlations are equal. There are, however, only two comparisons for which the difference exceeds .022. As it happens, both of the comparisons that yield relatively large differences involve quantitative tests. For Problem Solving, the difference is .066, and for Data Sufficiency, it is .084. These results suggest the possibility that quantitative tests given under different directions should not be regarded as strictly parallel. The limited data of this study, however, do not permit a firm conclusion on this point.

Effects of Directions on Score Equating: Method of Analysis

Two main approaches were used in determining the effect of differences in directions on score equating. In the first approach, each of the five operational parts was equated to itself. In the second approach, each of the five experimental tests given under Rights directions was equated to the corresponding experimental test given under Formula directions.

The first approach called for equating scores on an operational test to scores on the same operational test by the following three methods:

1. Identity Method. When a test is equated to itself, the ideal equating line has, by definition, a slope of 1 and an intercept of 0 and provides a standard with which results of other methods may be compared.

2. Invariant Link Method. In this method, each group takes one of the two tests that are to be equated. In addition, both groups take the same link test items, but under different guessing directions. One group takes the link test under Rights directions and the other group takes the link test under Formula directions. Equating is performed by rescoring by Formula the link test taken under Rights directions and assuming that

such scores can be treated as interchangeable with Formula scores earned under Formula directions. The analytical method used for treating these data is then identical to that for Maximum Likelihood equating, described below.

In this part of the study an operational test section was equated to the same operational test section as though two different tests were involved, using the data of the two spiralled groups that took the same experimental section (but under different directions). The experimental test given under Rights directions was rescored by Formula and used as the link test, as described in the preceding paragraph.

3. Spiralling Method. This method calls for distributing the tests in sequence within each room in which the test is administered. As a result of this process, the samples of students taking each form will represent systematic samples of the total group tested. According to probability theory, each subsample will tend to become increasingly similar to the other subsamples as sample sizes increase. Thus, for large samples it can be assumed that any two subsamples are approximately equal in the abilities measured by the tests to be equated. Scores on two tests are equated by setting equal the means and standard deviations of the samples taking those two tests. The result of the equating is that transformed scores on one test will have the same mean and standard deviation as the observed scores on the other test. (For a fuller discussion of this method, see Angoff (1971, pp. 569-571).)

Here, an operational test section was again equated, as it was by the Invariant Link Method, to the same operational test section as though two different tests were involved, using only the data of the two spiralled groups of students, as described in the preceding paragraph, and without the use of the experimental test scores as a link.

In this first approach, the primary interest was in comparing the results obtained using the Invariant Link Method with those obtained by the Identity and Spiralling methods.

Although it was not possible to express the results of these equatings on the customary GMAT scale, it was decided to establish an arbitrary scale for each part score, so defined that the mean converted score would be 500 and the standard deviation of converted scores would be 100 for the total study sample. In this way, equating results would be expressed on a scale similar to the GMAT Total score scale.

The second main approach called for equating Rights scores on an experimental test administered under Rights directions to Formula scores on the same experimental test, administered under Formula directions. Two methods of equating were used:

1. Maximum Likelihood Method. This method calls for administering each of the two tests to be equated to a random sample of a suitable group of students and administering the same link, or anchor, test to all members of both samples. In this study, the operational part corresponding to each pair of experimental tests served as the link test. The analytical procedure calls for the estimation of the mean and variance of both tests for the total combined sample, and for setting equal the

estimated means and standard deviations for the two tests, as is done in the Spiralling Method. The link test serves to increase the precision of the equating results. This method is described fully by Angoff (1971, pp. 576-579).

2. Invariance Method. In this method, the equating is based on the results of a single test administered to a single group. A test given under Rights directions and scored Rights is also scored by Formula. It is then assumed that the Formula scores so obtained are equivalent to the Formula scores that would have been obtained had Formula directions as well as Formula scoring been employed for that group. The equating procedure then calls for the direct equating of Rights scores to Formula scores for the same individuals by setting equal their means and standard deviations on the two types of scores. This procedure was carried out for the experimental tests using the data, in each instance, for the group taking the test under Rights directions.

This phase of the analysis made it possible to compare results obtained by the Invariance Method with those obtained by the Maximum Likelihood Method, which is a standard equating method.

In order to express the results of these equatings on a scale similar to the GMAT Total score scale, it was decided to equate Formula scores on each experimental test to Formula scores on the corresponding operational part, and to use these equations in conjunction with equations already developed relating Formula scores on each part to the arbitrary scale. The equating of experimental tests to the corresponding operational parts was done by setting means and standard deviations equal for examinees who took both tests. Algebraic solution of each pair of equations yielded equations relating Formula scores on the experimental tests to the arbitrary scale for each part. These results, when used along with the equations relating Rights scores on the experimental tests to Formula scores on the experimental tests made it possible to write equations to convert Rights scores on the experimental tests to converted scores in the units of the arbitrary scales.

Effect of Directions on Score Equating: Findings

Results obtained for equating each operational part to itself are shown in Table 30. In this table, the Identity and Spiralling methods yield results that do not involve the Invariance Hypothesis; the Invariant Link Method, however, does involve this hypothesis.

If the Identity Method results are taken as the standard of comparison, consideration of the slope values shows that the Invariant Link Method agrees more closely with the Identity Method than does the Spiralling Method in four of the five comparisons. For the 15 sets of results at selected points on the raw score scale, the Invariant Link agrees more closely with the Identity Method in seven comparisons, the Spiralling Method agrees more closely in four comparisons, and there are four ties. There is a marked similarity between the results of the Spiralling and Invariant Link methods for selected points. For mean scores, only one difference between the Spiralling and the Invariant Link results is as large as two converted score points, and in the remaining ten comparisons, only one difference is as large as three converted score points. It should be noted,

Table 30

Conversion Parameters Relating Formula Scores on Each Part of Operational GMAT to Formula Scores on Same Part as Determined by Various Methods of Equating[a]

Part of GMAT	Number of Items	Equating Method	Parameters: Slope	Parameters: Intercept	Converted Score When Raw (Formula) Score is: Chance	Mean[b]	Perfect
Reading Comprehension	25	Identity	18.6644	250.0109	250	500	717
Reading Comprehension	25	Spiral	18.7708	250.2330	250	502	720
Reading Comprehension	25	Inv.Link	18.6588	251.5824	252	501	718
Problem Solving	30	Identity	21.5332	311.4144	311	500	957
Problem Solving	30	Spiral	21.8347	311.5242	312	503	967
Problem Solving	30	Inv.Link	22.0155	309.8210	310	503	970
Practical Business Judgment	40	Identity	14.6552	222.2048	222	500	808
Practical Business Judgment	40	Spiral	14.6874	222.9918	223	501	810
Practical Business Judgment	40	Inv.Link	14.6669	221.2698	221	499	808
Data Sufficiency	30	Identity	15.9482	323.0691	323	500	802
Data Sufficiency	30	Spiral	16.2528	318.5462	319	499	806
Data Sufficiency	30	Inv.Link	16.2241	317.5239	318	498	804
Usage	25	Identity	18.4257	294.1554	294	500	755
Usage	25	Spiral	18.1806	297.8571	298	501	752
Usage	25	Inv.Link	18.2304	297.4223	297	501	753

---

[a] Each part score was expressed on a scale defined to have a mean of 500 and a standard deviation of 100 for the total group (N=55,780).

[b] Computed using mean score of total group for each part, as follows: Reading Comprehension, 13.3939; Problem Solving, 8.7579; Practical Business Judgment, 18.9554; Data Sufficiency, 11.0941; and Usage, 11.1716.

however, that both the Spiralling and the Invariant Link methods differ substantially from the Identity Method for perfect scores on the Problem Solving test. The results of this analysis may be interpreted as favorable to the usefulness of the Invariant Link Method under the conditions of the study.

Results shown in Table 31 permit a comparison of the Invariance Method with the Maximum Likelihood Method. The question of special interest is whether there is evidence of systematic differences in results between the two methods. With respect to slope parameters, the Maximum Likelihood Method yields a larger value in two of the five comparisons, the Invariance Method yields a larger value in two comparisons, and in the fifth comparison, the slopes are equal. Results for the selected raw score levels show a higher converted score for Maximum Likelihood in seven instances, a higher converted score for Invariance in seven instances, and one tie. Among the 15 comparisons only one difference exceeds four converted score points. For perfect scores on the Problem Solving test the Invariance Method yields a value 11 points higher than the Maximum Likelihood Method. These results are consistent with the other equating results in supporting that the hypothesis that Formula scores may be considered to be invariant with respect to Rights and Formula directions.

Table 31

Conversion Parameters Relating Rights Scores on Each Experimental Test to Formula Scores on the Same Experimental Test[a]

			Parameters		Converted Score When Raw (Rights) Score is:		
Experimental Test	Number of Items	Equating Method	Slope	Intercept	Chance	Mean[b]	Perfect
Reading Comprehension	29	Max. Lik.	16.1317	257.9860	352	500	726
Reading Comprehension	29	Invariance	15.9962	260.9961	354	501	725
Problem Solving	25	Max. Lik.	24.8160	259.6898	384	500	880
Problem Solving	25	Invariance	25.3876	255.9344	383	501	891
Practical Business Judgment	32	Max. Lik.	20.6554	110.4282	243	499	771
Practical Business Judgment	32	Invariance	20.6043	109.0225	241	497	768
Data Sufficiency	40	Max. Lik.	18.7842	193.7417	344	501	945
Data Sufficiency	40	Invariance	18.8972	189.2450	340	499	945
Sentence Correction	30	Max. Lik.	19.1585	172.4707	287	499	747
Sentence Correction	30	Invariance	19.1585	173.2982	288	500	748

[a] Each operational part score was expressed on a scale having a mean of 500 and a standard deviation of 100 for the total group (N=55,780). Formula Scores on each experimental test were equated to Formula Scores on the corresponding operational part by setting means and standard deviations equal for examinees who took both tests.

[b] Computed using mean Rights Score on experimental test for students who received Rights directions, as shown in Table 16.

Summary and Conclusions

Unlike the data provided by the first phase of the study, in which operational forms of the SAT-verbal and the Chemistry Achievement Test were administered in a specially designed experiment, the GMAT data were taken from a regularly scheduled administration of the test. The entire group of about 55,000 examinees who took the operational form of the GMAT in October 1980 were divided, essentially at random, into 10 approximately equal subgroups and assigned to take, in addition to the operational test, one of five available sections of pretest items--Reading Comprehension, Problem Solving, Practical Business Judgment, Data Sufficiency, and Sentence Correction--under either Rights or Formula directions. Except for Sentence Correction, there was an operational section representing the same item type. For Sentence Correction, the corresponding operational section included Usage rather than Sentence Correction items. The spiralled administration of these sections made it possible to compare responses of examinees made under the two types of directions and also to compare the characteristics of both Rights scores and Formula scores for the two types of directions.

The data provided by the GMAT phase of the study confirm the conclusion, drawn from the SAT data, that the response strategies of examinees are generally consistent with the instructions they are given for guessing. As evidenced by counts of items Omitted, Not Reached, and Guessed, examinees do attempt more items under Rights directions than under Formula directions. The result of this differential behavior is that they do, as expected, earn higher Rights scores under Rights directions than under Formula directions. However, when their answer sheets are rescored by Formula, it is found that the differences for examinees taking the tests under the two directions are virtually zero. This finding gives clear support to the Invariance Hypothesis, which is that Formula-scoring compensates for differences in guessing strategies caused by differences in directions. One interpretation of this finding is that although some students may indeed improve their scores by guessing on the basis of partial knowledge, other students appear to diminish their scores because they guess on the basis of misinformation. On the average, however, contrary to the Differential Effects Hypothesis, the guesses of all students taken together appear to be no better than chance. Also, as expected, examinees at lower ability levels show larger numbers of Omitted and Not Reached items than higher-ability examinees. However, contrary to the results found in the SAT study, the difference between the effects of the two sets of directions was smaller for high-ability students than for low-ability students.

The data of this phase of the study point to higher KR (20) reliabilities for Rights scoring, although there is a possibility that individual predilections to leave items unanswered may tend to inflate the reliability coefficients. This question regarding the interpretation of the reliability data will also bear somewhat on the question of the parallelism of Rights tests and Formula tests. In any case, it appears that two types of administration and scoring are not so different as to cause doubt regarding parallelism, at least in the case of the verbal subtests. Data for the quantitative subtests are less clear; questions of parallelism may well need closer scrutiny for quantitative types of items.

As would be anticipated from the results of the examination of the two opposing hypotheses, Invariance vs Differential Effects, the methods of equating that make use of the Invariance Hypothesis are in excellent agreement with those that are taken as criterion methods. These results are highly encouraging with respect to future attempts to equate Rights-administered-and-scored tests to Formula-administered-and-scored tests.

IMPLICATIONS OF FINDINGS

The data provided by the studies of College Board SAT-verbal, College Board Chemistry, and GMAT have helped considerably to clarify several of the issues relating to methods of administration and scoring of standardized tests. As described in the early pages of both parts of this report, the studies undertaken here were designed with several purposes in mind. Principal among these was the question whether Rights-administered-and-scored tests could be equated to Formula-administered-and-scored tests without endangering the continuity of meaning of the scale. But in the process of considering various methods for carrying out such equating operationally and for developing other equating methods and conversion equations as criteria for evaluating possible operational methods, it became clear that an assumption basic to these methods had to be satisfied. This was the assumption that on the [illegible]e, students respond by guessing, under Rights directions, to items that they would normally omit when confronted with the penalty for guessing imposed on them in Formula scoring, an assumption formally stated by Lord (1978). Therefore, granted this assumption, it was plain that although Rights scores earned under the two types of directions would be markedly different, Formula scoring would tend to obliterate these differences.

The foregoing assumption, which is the basis of the Invariance Hypothesis, was supported by the data in both studies reported here, the College Board studies and the GMAT study. As expected, the

Rights scores of examinees tested under Rights directions were much higher than the Rights scores of examinees tested under Formula directions. But when the answer sheets for the two groups of examinees were rescored by Formula, the differences between the groups virtually disappeared. Moreover, studies of SAT-verbal examinees indicate that this finding applies not only overall, but also separately at different levels of ability. Thus, it is not true, as might have been expected, that students at some levels of ability are more perceptive regarding their assessments of their own knowledge than students at other levels of ability. Apparently, students at all levels of ability are equally unable to discern differences in their own levels of competence at the edges of their competencies. Guessing at those edges appears to be as much influenced by misinformation as by valid information.

The fact that the Invariance Hypothesis is supported, not only overall, but for examinees at all levels of ability, is of considerable importance for at least three reasons: One, it disconfirms the assertion made in the Differential Effects Hypothesis, which is that students are disadvantaged by Formula directions, that they would be better advised to guess, even in a Formula-scored test, since their scores would be higher, on the average, than if they did not guess. The fact is, however, that their scores would not be higher if they guessed than if they did not guess. Moreover, it seems to be assumed by the proponents of the Differential Effects Hypothesis that Rights directions equalize the advantage for all students, because Rights directions encourage students to respond to all the items. However, as we have observed in this study, the numbers of items Omitted and Not Reached, although smaller in Rights-directed than in Formula-directed tests, are still substantial; contrary to hypothesis, students do not all respond to all the items, in spite of the strong directions. Two, and most central to the particular purpose of this investigation, the evidence for the Invariance Hypothesis makes it possible to equate Rights scores to Formula scores without experiencing unacceptably large slippage in the scale, even under conditions of test disclosure, were our programs to change from Formula-scoring to Rights-scoring. As the studies of equating Rights to Formula indicate, the use of the Invariance Hypothesis makes the transition entirely feasible. Three, it is important to observe that the confirmation of the Invariance Hypothesis implies that since Formula scoring has the effect of compensating for, or equalizing, differences in behavior resulting from different directions for guessing, it also has the effect of compensating for differences in individual student strategies for guessing. Not only does this property of Formula-scoring have significance for easing the transition from Formula-scored tests to Rights-scored tests, it also has a more basic significance for the test administration itself.

The tabulations of the nonresponse data confirm the findings made in other analyses of the data: Nonresponse is a function of the directions given in the administration, and also a function of ability level, but not, at least as evidenced in these data, a function of ethnicity. There are fewer items Omitted, Not Reached, and Not Attempted, and, correspondingly, more items Guessed (as measured by either the W-0 or the Ziller index) for students tested under Rights directions than for students tested under Formula directions. Also, there are fewer items Omitted, Not Reached, and Not Attempted by more able than by less able students. To what extent this finding is a function of ability in the abstract sense and to what extent it is a function of the constraint imposed on the scores by the number of items in the test is difficult to know. Concerning whether abler students respond more appropriately to directions for guessing than less able students, the data of the two studies yielded inconsistent results.

The fact that examinees answered more items under Rights directions than under Formula directions is in accordance with expectations. However, the expectation that every examinee would answer every item under Rights directions was by no means fulfilled, despite the fact that the instructions stated explicitly that it would be to their advantage to do so. These results emphasize the importance of systematic efforts to encourage examinees to answer every question if Rights-directions-and-scoring are adopted for operational testing. Indeed, under operational conditions, a determined effort to minimize the number of unanswered items may be considered to be an important step in maintaining uniform testing conditions for all examinees,

Whether guessing is a function of ability level is difficult to say. This, it appears, depends on the operational definition of guessing one is willing to accept. As was pointed out earlier in this report, the tendency to guess, when conceived of abstractly as a personality trait, is probably uncorrelated with cognitive ability; on the other hand, guessing behavior, certainly when derived from the test responses themselves, would necessarily be correlated with test score. Whether guessing behavior is better expressed as the index, W-O, or as a proportion of noncorrect responses, as in the Ziller index, or in some index other than either of these, is indeterminate. Yet it is basic to our conclusions because in one respect, at least, the two indices lead to different conclusions: the W-O index is negatively correlated with test scores; the Ziller index is positively correlated with test scores, but in general, the absolute size of the coefficients is smaller.

Although the parallel-forms reliabilities are virtually equal to the two types of administration and scoring, the KR (20) reliability coefficients are not: the reliabilities for Rights-administered-and-scored tests have a small, but consistent edge over the reliabilities for the Formula tests. Here too, however, the interpretation is not entirely clear. If there are consistent differences among individuals with respect to the tendency to guess, such differences will inevitably become confounded with the scores themselves, but in such a way as to inflate the reliability coefficients, however they are calculated; and until guessing as a personality trait can be reliably measured and shown to correlate more with one type of administration than the other, this question too must remain indeterminate.

The point has often been made that the issue of Rights vs Formula is a trivial one because the two scores are so highly correlated; given a set of answer sheets, the correlation between the two scores is usually in excess of .98 or even .99. Quite aside from the appropriateness of the conclusion of trivialness, the evidence in support of it is clearly spurious inasmuch as both scores are based on the same set of test responses and therefore must perforce by highly correlated. An appropriate way to evaluate this question, it is submitted, is to assemble data of the sort designed in the study of SAT-verbal, in which randomly different groups take the same pair of tests under the two conditions of administration. These data make it clear that in fact two tests administered and scored in the same way, Rights or Formula, correlate more highly than two tests administered and scored in different ways. However, the differences amount to only about .02, on the average, when the correlations are in the vicinity of .80.

Closely related to the foregoing question is the question of parallelism of the Rights-administered-and-scoring mode vs the Formula-administered-and-scoring mode. The data from the SAT-verbal study are clear on this point: Although it is true that true-score correlations between tests that are administered and scored in the same mode are higher than true-score correlations between tests that are administered and scored in different modes, the differences are small. In any case, the true-score correlations between Rights and Formula tests are close enough to unity to dispel any concerns that the two types of administration and scoring are measuring different abilities.

The data from the GMAT study are less clear on this point. The results of the verbal tests are essentially in agreement with the SAT data, differing from the latter chiefly in the respect that some of the GMAT subtests may be less homogeneous and therefore less reliable in the KR (20) sense than the SAT subtests. The differences observed in the case of the GMAT quantitative tests, however, are somewhat larger. The assumption of parallelism for such tests may not be fully warranted.

The implications of the findings of the College Board and GMAT studies for the success of equating efforts in effecting a change from Formula-type tests to Rights-type tests are on the whole quite positive. The methods of equating that have been examined here for possible use in operational equating work have made use of the Invariance Hypothesis, and, as expected from the earlier confirmation of this hypothesis, these methods yield results that are in good agreement with other, more nearly ideal procedures. Even if these results fall short of expectations, the data of the study have made it clear that students can and do shift their mode of response to test items in accordance with changes in directions for guessing, and moreover, appear to do so even in operational test administrations, when they might be expected to perceive that a particular test section is experimental and will not count toward their score. Supported by evidence of this sort, and supported further by the results of these studies that show that differences in guessing strategies tend to be overcome and removed by Formula scoring, we may feel encouraged that still other methods of equating may be developed to supplement those examined in this study to enlarge the range of possible solutions to the problems of equating across a transition.

Beyond the purposes for which this study was designed, and the insights it has permitted into a set of issues that have so long been the subject of controversy, some mention should be made of the value of the type of experimental design used in this study, one that is likely, if adopted by other investigators in the future, to clarify still other issues yet unresolved. As was pointed out in the early pages of this report, the present study of Rights and Formula scoring is the only one to our knowledge that has been based on very large samples and designed in a symmetrical fashion, with an essentially random half of the examinees exposed to one type of directions for guessing and the other half exposed to another type of directions. This arrangement, supported, when possible, with additional, relevant test scores administered in the same way to everyone, as was the case in this study, and with background data--age, sex, and ethnic membership, for example--would serve to enhance the informational quality of future studies.

As a result of the random assignment of large groups of students to the two types of directions and the use of ability and background controls, these two sets of data--the SAT-verbal and Chemistry Achievement Test data, and the GMAT data--have considerable value for other studies of the effects of test administration and scoring. These could involve, for example, studies of speededness under the two conditions of administration (some of which have already been done), modifications of the W-O and the Ziller indices of guessing, more detailed examination of the Invariance Hypothesis in the GMAT administration as a function of ability level, studies of scoring accuracy, studies of parameter estimation for important applications of item response theory, studies of conventional methods of equating, and undoubtedly many others. Such studies could be undertaken and carried out to great advantage without the substantial costs of special administration costs which often tend to inhibit the conduct of potentially useful studies.

References

Abu-Sayf, F. K. The scoring of multiple-choice tests: A closer look. Educational Technology, 1979, 19, 5-15.

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.) Educational Measurement. Washington, D.C.: American Council on Education, 1971.

Boldt, R. F. Study of linearity and homoscedasticity of test scores in the chance score range. Educational and Psychological Measurement, 1968, 28, 47-60.

Cross, L. H. and Frary, R. B. An empirical test of Lord's theoretical results regarding formula scoring of multiple-choice tests. Journal of Educational Measurement, 1977, 14, 313-321.

Cureton, E. E. The correction for guessing. Journal of Experimental Education, 1966, 34 (4), 44-47.

Davis, F. B. A note on the correction for chance success. Journal of Experimental Education, 1967, 35 (3), 42-47.

Diamond, J. and Evans, W. The correction for guessing. Review of Educational Research, 1973, 43, 181-191.

Dressel, P. L. Some remarks on the Kuder-Richardson reliability coefficient. Psychometrika, 1940, 5, 305-310.

Ebel, R. L. Measuring educational achievement. Englewood Cliffs, N.J.: Prentice-Hall, 1965.

Ebel, R. L. Blind guessing on objective achievement tests. Journal of Educational Measurement, 1968, 5, 321-325.

Glass, G. V and Wiley, D. E. Formula scoring and test reliability. *Journal of Educational Measurement*, 1964, *1*, 43-47.

Levine, R. and Lord, F. M. An index of the discriminating power of a test at different parts of the score range. *Educational and Psychological Measurement*, 1959, *19*, 497-503.

Lord, F. M. Formula scoring and validity. *Educational and Psychological Measurement*, 1963, *23*, 663-672.

Lord, F. M. *Relative efficiency of number-right and formula scores*. Research Bulletin 74-9. Princeton, N.J.: Educational Testing Service, 1974.

Lord, F. M. Formula scoring and number-right scoring. *Journal of Educational Measurement*, 1975, *12*, 7-11.

Lord, F. M. Practical applications of item characteristic curve theory. *Journal of Educational Measurement*, 1977, *14*, 117-138.

Lord, F. M. *Applications of item response theory to practical testing problems*. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1980.

Lord. F. M. and Novick, M. R. *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley, 1968.

Rowley, G. L. and Traub, R. E. Formula scoring, number-right scoring, and test-taking strategy. *Journal of Educational Measurement*, 1977, *14*, 15-22.

Sherriffs, A. C. and Boomer, D. S. Who is penalized by the penalty for guessing? *Journal of Educational Psychology*, 1954, 45, 81-90.

Slakter, M. J. Risk taking on objective examinations. *American Educational Research Journal*, 1967, *4*, 31-43.

Slakter, M. J. The penalty for not guessing. Journal of Educational Measurement, 1968, 5, 141-144(a).

Slakter, M. J. The effect of guessing strategy on objective test scores. Journal of Educational Measurement, 1968, 5, 217-226(b).

Slakter, M. J. Generality of risk taking on objective examinations. Educational and Psychological Measurement, 1969, 29, 115-128.

Stanley, J. C. Psychological correction for chance. Journal of Experimental Education, 1954, 22, 297-298.

Swineford, F. The measurement of a personality trait. Journal of Educational Psychology, 1938, 29, 295-300.

Swineford, F. Analysis of a personality trait. Journal of Educational Psychology, 1941, 32, 438-444.

Thorndike, R. L. (Ed.) Educational Measurement. Washington, D.C.: American Council on Education, 1971.

Votaw, D. F. The effect of do-not-guess directions upon the validity of true-false or multiple-choice tests. Journal of Educational Psychology, 1936, 27, 698-703.

Ziller, R. C. A measure of the gambling response-set in objective tests. Psychometrika, 1957, 22, 289-292.